

On the origin of thermality

Bernard S. Kay*

Department of Mathematics, University of York, York YO10 5DD, UK

It is well-known that a small system weakly coupled to a large energy bath will, when the total system is in a microcanonical ensemble, find itself to be in an (approximately) thermal state (i.e. canonical ensemble) and, recently, it has been shown that, if the total state is, instead, a random pure state with energy in a narrow range, then the small system will still be approximately thermal with a high probability (defined by ‘Haar measure’ on the total Hilbert space). Here we ask what conditions are required for something resembling either/both of these ‘traditional’ and ‘modern’ thermality results to still hold when the system and energy bath are *of comparable size*. In Part 1, we show that, for given system and energy-bath densities of states, $\sigma_S(\epsilon)$ and $\sigma_B(\epsilon)$, thermality does not hold in general, as we illustrate when σ_S and σ_B both increase as powers of energy, but that it does hold in certain approximate senses, in both traditional and modern frameworks, when σ_S and σ_B both grow as $e^{b\epsilon}$ or as $e^{q\epsilon^2}$ (for constants b and q) and we calculate the system entropy in these cases. In their ‘modern’ version, our results rely on new quantities, which we introduce and call the S and B ‘modapprox’ density operators, which are defined for any positively supported, monotonically increasing, σ_S and σ_B , and which, we claim, will, with high probability, closely approximate the reduced density operators for the system and energy bath when the total state of system plus energy bath is a random pure state with energy in a narrow range. In Part 2 we clarify the meaning of these modapprox density operators and give arguments for our claim.

The prime examples of non-small thermal systems are quantum black holes. Here and in two companion papers, we argue that current string-theoretic derivations of black hole entropy and thermal properties are incomplete and, on the question of information loss, inconclusive. However, we argue that these deficiencies are remedied with a modified scenario which relies on the modern strand of our methods and results here and is based on our previous *matter-gravity entanglement hypothesis*.

PACS numbers: 03.65.Yz, 05.30.Ch, 04.70.Dy, 04.60.Cf

I. INTRODUCTION

A. Background

This paper is concerned with the general question: “How do physical systems get to be hot?”. By ‘hot’ here, we do not simply mean ‘having lots of energy’. We shall reserve the word ‘energetic’ for that. Rather, we mean the more specialized notion of being in what is known, in (quantum) statistical mechanics, as a Gibbs state, i.e. a state described by a density operator of form

$$\rho_\beta^{\text{Gibbs}} = Z_\beta^{-1} e^{-\beta H} \quad (1)$$

where H is a suitable (usually, of necessity, approximate) Hamiltonian (assumed to have discrete spectrum) for the system and β is related to the system’s temperature, T , by $\beta = 1/kT$ where k is Boltzmann’s constant (henceforth set to 1). Here Z_β stands for $\text{tr}(e^{-\beta H})$ and is the normalization constant which ensures that $\rho_\beta^{\text{Gibbs}}$ will have unit trace. (When regarded as a function of β it is, of course, the system’s ‘partition function’.) Such states are also known as ‘canonical states’ or ‘thermal equilibrium states’ or ‘KMS states’. We shall sometimes refer to them simply as ‘thermal’ states. A possible source of confusion here is the fact that it is sometimes found to be convenient to adopt the fiction that a system which is merely energetic is in a Gibbs state at a temperature chosen so as to give it the same mean energy. Additionally, given a system with a density of states $\sigma(\epsilon)$, it can sometimes be convenient to assign to it a ‘temperature’, $T(\epsilon)$, at each energy, ϵ , according to the formula $1/T(\epsilon) = d \log \sigma(\epsilon)/d\epsilon$ [1]. We wish to underline that we shall not be concerned with such a fiction, nor with such an assignment of an energy-dependent ‘temperature’, here. Rather we are interested in how systems get into states which are *actually* Gibbs states. In particular, we are interested in black bodies, and, more particularly, black holes (in suitable boxes; here we refer

*Electronic address: bernard.kay@york.ac.uk

to the remarkable developments in ‘Euclidean Quantum Gravity’ and in (Quantum) ‘Black Hole Thermodynamics’ which arose from Hawking’s pioneering work [2] on ‘Black Hole Evaporation’ – see e.g. the papers on quantum black holes in the collections [3, 4]).

Of course, one way for a system to get into a Gibbs state is for it to be weakly coupled to a (much larger) heat bath which is already in a Gibbs state at the desired temperature. There is a considerable literature, which, with varying degrees of mathematical rigour and generality, shows that, as one might expect, a typical such system will, more or less irrespective of its initial state, approximately get into a Gibbs state at the same temperature at late times – see e.g. [5], [6]. However, what we are really interested in when we ask our general question “How do physical systems get to be hot?” is:

“How does any physical system ever get to be hot in the first place?”

Obviously, an explanation of how one system gets to be hot which invokes the existence of another system (the above-mentioned heat bath) which is assumed already to be hot can’t help to answer this version of our question!

Another traditional explanation for the propensity of some systems to be in Gibbs states goes along the following lines (see e.g. [7] and, for a treatment of some of the related mathematical aspects, e.g. [8] as well as the paper [9] which recalls this traditional explanation as a preliminary to its main purpose – for which see below): One assumes one’s system of interest, say described by a Hamiltonian, H_S , on a Hilbert space, \mathcal{H}_S , to be weakly coupled to a much larger ‘energy bath’, with Hamiltonian, H_B , on a Hilbert space, \mathcal{H}_B – both Hamiltonians being assumed to have a finite number of energy levels in any finite energy interval, with the number of states of the energy bath in an energy interval, δ , being approximately given in terms of a ‘density of states’, σ_B , as $\sigma_B(\epsilon)\delta - \sigma_B$ being assumed to have some typical, say, power-law form (see below) – and one assumes the whole system to be in a total microcanonical state. Before we explain what we mean by this, we pause to remark, first, that, in order to avoid ambiguous usages of the word ‘system’, we shall, from now on, adopt the word *totem* (short for ‘total system’) to denote what we referred to above as our ‘whole system’. So we shall talk about a ‘totem’ which consists of a ‘system’. ‘S’, and an ‘energy bath’, ‘B’. Our assumption of weak coupling is then the assumption that the totem Hamiltonian will take the form

$$H = H_S \otimes 1 + 1 \otimes H_B + \text{weak coupling term} \quad (2)$$

on the totem Hilbert space, $\mathcal{H} = \mathcal{H}_S \otimes \mathcal{H}_B$, and we shall assume further that the coupling term is so weak that it can be neglected for state and energy-level counting purposes. To say that our totem is in a microcanonical state then means to assume it is described by the density operator

$$\rho_{\text{microc}} = M^{-1} \sum |\epsilon\rangle\langle\epsilon| \quad (3)$$

on the totem Hilbert space, \mathcal{H} , where the sum is over a basis of energy eigenstates for the subspace of \mathcal{H} consisting of energy levels with energies in an interval, $[E, E + \Delta]$, which is small, yet large enough for the total number of totem energy eigenstates in this range to be very large, while the normalization constant, M (which is expected to roughly scale with Δ) is the total number of such basis eigenstates. We further pause to note that we shall assume throughout the present paper, as is usually assumed for ‘ordinary’ physical systems, that both Hamiltonians, H_S and H_B , are positive and their densities of states monotonically increasing. We remark though that, as we will discuss further in Section VIII, were any of these assumptions to be relaxed, then the prospects for systems to become hot become much less constrained and, in particular, there are ways in which a system can be hot while the totem is in a pure state which differ from the ‘modern’ scenarios we discuss below.

Proceeding with the above assumptions, the states, $|\epsilon\rangle$, in the sum in (3) will each take the form $|\epsilon_S\rangle \otimes |\epsilon_B\rangle$ and the sum over totem energy levels will become (see (6) below) a double sum over system energy levels, ϵ_S , and energy-bath energy levels, ϵ_B , which satisfy the condition $\epsilon_S + \epsilon_B \in [E, E + \Delta]$. The resulting state of the system is then represented mathematically, in the usual way, by the reduced density operator, ρ_S^{microc} on \mathcal{H}_S i.e. by the partial trace of ρ_{microc} over \mathcal{H}_B .

To remind ourselves how thermality of our system can then come about in this traditional explanation, it is instructive first to consider an oversimplified model in which our system Hilbert space, \mathcal{H}_S , is two-dimensional with only two energy levels with energies ϵ_S^1 and ϵ_S^2 such that $\epsilon_S^2 - \epsilon_S^1 \gg \Delta$ and in which the density of states, σ_B , of the energy bath grows exponentially – we shall write $\sigma_B(\epsilon) = ce^{b\epsilon}$. (We shall discuss the case where both system and energy bath both have such a density of states in Sections III and V.)

Then we easily see that ρ_S^{microc} will be approximately

$$\rho_S^{\text{microc}} = n^{-1}(c\Delta e^{b(E-\epsilon_1)}|\epsilon_S^1\rangle\langle\epsilon_S^1| + c\Delta e^{b(E-\epsilon_2)}|\epsilon_S^2\rangle\langle\epsilon_S^2|)$$

where n denotes the appropriate normalization constant, and this is clearly the same as the Gibbs state

$$\rho_\beta = Z_\beta^{-1}(e^{-\beta\epsilon_1}|\epsilon_S^1\rangle\langle\epsilon_S^1| + e^{-\beta\epsilon_2}|\epsilon_S^2\rangle\langle\epsilon_S^2|)$$

for $\beta = b$ for a suitable, normalizing, Z_β .

In the full story, where we now assume that also the states of the system are approximately given by a density of states, σ_S , it is convenient to assume that E is an integral multiple of Δ and locally to slightly distort the spectra of system and energy bath so that their energy levels are evenly spaced at intervals $\Delta, 2\Delta, \dots, E$ with each system level having degeneracy

$$n_S(\epsilon) = \sigma_S(\epsilon)\Delta \quad (4)$$

and each energy-bath level having degeneracy

$$n_B(\epsilon) = \sigma_B(\epsilon)\Delta. \quad (5)$$

If, as we shall further assume, this can be done in such a way as to maintain the same ‘smoothed out’ densities of states, then it will not seriously alter the values of any quantities of interest. Choosing a basis within the degeneracy subspace of \mathcal{H}_S with each energy, ϵ , and labelling its elements $|\epsilon, i\rangle$, where, for each ϵ , $i = 1, \dots, n_S(\epsilon)$ while ϵ ranges from Δ to E in integer steps of Δ (and similarly for the energy bath) we then easily have that ρ_{microc} (3) can be rewritten as

$$\rho_{\text{microc}} = M^{-1} \sum_{\epsilon_S} \sum_{\epsilon_B} \sum_i \sum_j |\epsilon_S, i\rangle \otimes |\epsilon_B, j\rangle \langle \epsilon_S, i| \otimes \langle \epsilon_B, j| \quad (6)$$

where the sum over i goes from 1 to $n_S(\epsilon_S)$, the sum over j goes from 1 to $n_B(\epsilon_B)$ and the sums over ϵ_S and ϵ_B are over values which are positive-integer multiples of Δ and are constrained to have $\epsilon_S + \epsilon_B = E$, while the normalization constant, M , defined after (3), is also given by

$$M = \sum_{\epsilon=\Delta}^E n_S(\epsilon) n_B(E - \epsilon) \quad (7)$$

or, roughly equivalently [14], by making the replacement

$$\sum_{\epsilon=\Delta}^E \text{ by } \Delta^{-1} \int_0^E d\epsilon \quad (8)$$

by the approximate formula

$$M = \Delta \int_0^E \sigma_S(\epsilon) \sigma_B(E - \epsilon) d\epsilon. \quad (9)$$

Moreover, δ/Δ times the summand in (7) or $\delta\Delta$ times the integrand in (9) is, for suitable (small but not too small) δ (approximately) the number of energy eigenstates for which the energy of the totem lies in the interval $[E, E + \Delta]$ while the energy of the system lies in the interval $[\epsilon, \epsilon + \delta]$. When our totem is in the microcanonical state (3), (6), this summand divided by M may thus be interpreted as the probability that the system energy lies in this latter interval. We shall denote it by $P_S(\epsilon)\delta$ and call $P_S(\epsilon)$ the system’s *energy probability density* so we have

$$P_S(\epsilon) = \frac{\Delta}{M} \sigma_S(\epsilon) \sigma_B(E - \epsilon) \simeq \frac{1}{M\Delta} n_S(\epsilon) n_B(E - \epsilon), \quad (10)$$

and we notice, in passing, that

$$P_B(\epsilon) = P_S(E - \epsilon).$$

The reduced density operator, ρ_S^{microc} , of ρ_{microc} on \mathcal{H}_S will clearly be

$$\rho_S^{\text{microc}} = M^{-1} \sum_{\epsilon=\Delta}^E n_B(E - \epsilon) \sum_{i=1}^{n_S(\epsilon)} |\epsilon, i\rangle \langle \epsilon, i|. \quad (11)$$

(Here and below, to avoid cluttering up our formulae, we drop the ‘s’ suffix on ϵ – also in $|\epsilon, i\rangle$ – when there can be no ambiguity.)

One can then show, for a wide range of ‘realistic’ energy-bath models that, in the limit as the energy bath gets large while the system remains unchanged, ρ_S^{microc} will converge to a thermal state at an inverse temperature β given, [15], in terms of the large-size behaviour of the energy bath’s density of states.

In particular, and specializing now to a case (cf. again e.g. [7]) that will interest us further below, if the density of states, σ_B , has the typical power-law form of ordinary (radiationless) matter:

$$\sigma_B(\epsilon) = A_B \epsilon^{N_B}, \quad (12)$$

where A_B is a constant and N_B is an ‘Avogadro-sized’ number which could stand e.g. for ‘3/2 times the number of molecules’ in the energy bath or the ‘number of oscillators’ in the energy bath (see again e.g. [7] for the origin of the 3/2 etc.) etc. then, in the limit as the total energy, E , of the totem gets larger while the size of the energy bath gets larger – in the sense that N_B gets larger – while the system remains unaltered and N_B/E converges to a constant, β , ρ_S^{microc} will converge to a thermal state at inverse temperature β – i.e. to the $\rho_{S,\beta}^{\text{Gibbs}}$ of Equation (13) below. In the special case that the system has a density of states also of power-law form (see (18)) we shall provide a proof of this result, in passing, in Section II below which is particularly instructive in relation to our present purposes. See the last paragraph in Section II. So, in this way, one shows that a small system in contact with a large energy bath with a suitable density of states will approximately be in a Gibbs state when the totem is in a microcanonical state.

Above, a Gibbs state (1) of our system will obviously take the form (assuming again the spectrum to be slightly distorted as explained before equation (6))

$$\rho_{S,\beta}^{\text{Gibbs}} = Z_{S,\beta}^{-1} \sum_{\epsilon=\Delta}^{\infty} e^{-\beta\epsilon} \sum_{i=1}^{n_S(\epsilon)} |\epsilon, i\rangle \langle \epsilon, i| \quad (13)$$

where (approximating the obvious sum by an integral as we did when we passed from (7) to (9))

$$Z_{S,\beta} = \int_0^{\infty} \sigma_S(\epsilon) \exp(-\beta\epsilon) d\epsilon. \quad (14)$$

However, this traditional explanation of the origin of thermality (of a small system) is also unsatisfactory since it still begs the question of how the totem got into a microcanonical state. What would really be desirable would be an explanation of the origin of thermality consistent with the basic assumption of standard quantum mechanics that the total state of a closed system (in our case, our totem) is a pure state – i.e. in the language of density operators, the projector, $|\Psi\rangle\langle\Psi|$, onto a single vector, Ψ , in the closed system’s (/our totem’s) Hilbert space.

Such an explanation has, in fact, recently been given by a number of authors again for the case of a small system in contact with a large energy bath. See especially the paper [9] entitled ‘Canonical Typicality’ by Goldstein, Lebowitz et al. and also the references therein. The result of that paper – when specialized to our power-law density of states model (12) – amounts to the statement that if, for a ‘system’ and ‘energy bath’ as considered above, one takes a random pure state with energy in the energy range $[E, E + \Delta]$, then, again imagining the energy bath to get larger while N_B/E converges to β , for sufficiently large E , the reduced density operator of the system, ρ_S^{modern} , will, with very high probability, be very close to a Gibbs state (i.e. the $\rho_{S,\beta}^{\text{Gibbs}}$ of (13)) at inverse temperature β .

We shall also re-obtain this result ourselves as a limiting case of one of our main new results in Section ID.

The precise mathematical statement can be inferred by inspecting the paper [9] and/or see the more general rigorous result proved by Popescu et al. [12].

Goldstein, Lebowitz et al. define what they mean here by ‘random’ and by ‘probability’ by taking the natural measure on the set of unit vectors of the relevant Hilbert space – assumed to have large, but finite, dimension M – by thinking of it as a $(2M - 1)$ -dimensional real unit sphere and taking the natural invariant measure induced on that by Haar measure on the orthogonal group. In doing so, they follow pioneering work of Lubkin [10] who, in 1978, after introducing [11] this use of this measure (following Lubkin and subsequent authors, we shall simply call it ‘Haar’ measure from now on) showed that a randomly chosen pure density operator, $\rho^{mn} = |\Psi\rangle\langle\Psi|$ (without any restriction on energy or anything else) on the tensor-product Hilbert space, $\mathcal{H}_m \otimes \mathcal{H}_n$, of a pair of quantum systems – \mathcal{H}_m being m -dimensional and \mathcal{H}_n being n -dimensional – will, for fixed m and $n \gg m$, have, with high probability, a reduced density operator, ρ_m^{mn} , on \mathcal{H}_m , which is close to the maximally mixed density operator – with components, in any Hilbert space basis, $\text{diag}(1/m, \dots, 1/m)$. We shall discuss further this result of Lubkin and some related developments in Section X at the beginning of Part 2 since they will be needed as a preliminary towards our argument for Equation (15) and the related claimed proposition in Section ID.

In essence, one might characterize the relation between Lubkin’s work and the work, [9], of Goldstein, Lebowitz et al. by saying that Lubkin obtained microcanonicity of a small subsystem from randomness of a totem pure state while Goldstein, Lebowitz et al. obtained canonicity of a small subsystem when an, otherwise random, totem pure

state is constrained to have a definite energy. (Popescu et al. [12] then generalized these developments by allowing for more general constraints, and also made them mathematically rigorous.)

The modern (see Endnote [13]) results, [9, 12], of Goldstein, Lebowitz et al. and of Popescu et al. are an advance on the traditional results in that they replace the assumption of a total microcanonical state by the assumption of a total pure state. However, they still share the limitation of the traditional approach of still only being capable of explaining how, at most, only a small subsystem of a given ‘large’ totem can get to be (approximately) thermal. The main purpose of the present paper will be to explore to what extent, and/or under what altered circumstances, this limitation can be overcome. Our main motivation relates to the theory of quantum black holes. Black holes are a puzzle in relation to the above results if one believes, as seems compelling, that the totem consisting of a black hole in equilibrium with its atmosphere in a box at approximately fixed energy is completely (approximately) thermal [16].

B. Quantum black holes

In such black hole equilibrium states we may roughly (albeit not exactly, see Endnote (iii) in [19]) identify the black hole itself with ‘gravity’ and the atmosphere with ‘matter’. In an earlier proposal (see [17, 18] and especially Endnotes (i), (ii), (iii) and (v) in [19]) of the author (which predated the work [9, 12] in a more general, but non-gravitational [25], context of Goldstein, Lebowitz et al. and of Popescu et al. by around seven years) a radically-different-from-usual hypothesis was put forward as to the nature of quantum black hole equilibrium states according to which the total state is a pure state (in line with what we are calling here the ‘modern’ approach – see Endnote [13] – but in contrast to the usual assumption in work on quantum black holes that it is a Gibbs state at the Hawking temperature) while the reduced state of the gravitational field alone and also the reduced state of the matter fields alone are each thermal (i.e. each Gibbs states) at the appropriate Hawking temperature (see below). (Here we use the word ‘matter’ to include e.g. the electromagnetic field.) Below, we shall sometimes call such a total pure state *bithermal*. This hypothesis formed, in turn, just a part of our wider hypothesis [17–19] (which we shall sometimes refer to here as our *matter-gravity entanglement hypothesis*) according to which, quite generally, one should always take into account the quantum gravitational field as well as all matter fields in describing the full dynamics of any physically closed totem, and that, while the state of the totem is always pure and evolves unitarily, the ‘physically relevant’ quantum state is to be identified with the reduced density operator of the matter alone and, concomitantly (see Section I E and, in particular, Endnote [48]), the physical entropy of a closed totem is to be identified with its *matter-gravity entanglement entropy*. Interpreted according to this wider hypothesis, our hypothesis that quantum black hole equilibrium states are bithermal then implies that, *physically*, such states are *completely thermal*. We remark that, given our wider hypothesis, what is required for this complete thermality is, of course, just thermality of the reduced state of the matter. However, there are strong reasons (particularly the fact [4] that the Euclideanized Schwarzschild metric is periodic in imaginary time with period $8\pi G\mathcal{M}$) for believing that the mathematical nature of the reduced state of gravity will also be thermal and this is what we have assumed above and will continue to assume in the remainder of this section and in Section IX.

To summarize and also to recall the relevant formulae: While we accept the (conventional) belief that, in black hole equilibria, both matter and gravity are each separately thermal at the Hawking temperature, T_H , we propose (unconventionally in comparison to other work on quantum black holes) that the total state of matter-gravity is pure (rather than itself being a thermal state). The thermality of each of the reduced states (i.e. of matter and of gravity separately) will then arise as the result of entanglement between matter and gravity in the pure totem state. We shall refer to this picture of black hole equilibrium states as our *entanglement picture of black hole equilibrium*. (We shall assume in Section IX and in [22, 23] that, in this picture, the overall (i.e. totem) state of black hole equilibrium is not only pure but also close to an energy eigenstate.) We further emphasize that while this proposal is unconventional when compared to other work on quantum black holes, it seems to fit well with modern approaches (such as those of [9, 12]) towards understanding the origin of thermality which have recently been proposed in non-gravitational contexts. Here, we recall that the Hawking temperature, T_H , is given [3, 4], in the case of a Schwarzschild (i.e. spherical, uncharged) black hole of mass \mathcal{M} , by $T_H = 1/8\pi G\mathcal{M}$ (in general the surface gravity multiplied by 2π). Here, G denotes Newton’s constant and we set c and \hbar to 1. Moreover, we accept the conventional belief that the physical entropy – again in the spherical, uncharged case – has the Hawking value of $4\pi G\mathcal{M}^2$ (in general, one quarter of the area of the event horizon, divided by G) and what is new about our proposal is our claim that this entropy-value should ultimately be explainable as the matter-gravity entanglement entropy of a pure state of the overall matter-gravity totem.

Finally, we note that our matter gravity entanglement hypothesis and our entanglement picture of black hole equilibrium also offer a natural resolution to the Information Loss Puzzle [20]. This puzzle arose because, as long as it was believed that black holes were correctly described by mixed states, then, in a dynamical process in which black holes were formed from collapsing stars etc., it appeared that an initial pure state would dynamically evolve

into a mixed state, contradicting unitarity. On the other hand, there is no difficulty in reconciling our matter-gravity entanglement hypothesis with a unitary quantum mechanical time evolution and, once we identify entropy as matter-gravity entanglement entropy, this is entirely consistent with increasing entropy (i.e. information loss). We note that this proposed resolution to the Information Loss Puzzle is, in fact, just a special case of our proposed resolution to the Second Law Puzzle [17, 19, 22].

C. Our specific question

The specific question we shall endeavour to answer in this paper assumes, as its basic setting, that a totem be given which consists of a pair of weakly coupled systems, S and B, each with its own Hilbert space, \mathcal{H}_S and \mathcal{H}_B , and each with its own density of states, σ_S and σ_B .

Our specific question is then:

If the systems, S and B, are of comparable size [24], what modifications need to be made either to the traditional ‘total microcanonical state’ approach or, more relevantly since we believe it to be a step closer to the right answer, to the more modern ‘total pure state’ approach of Goldstein, Lebowitz et al. and of Popescu et al. and others, as described above, so as to ensure that when the totem has a total state with energy in an interval $[E, E + \Delta]$, the reduced states of S and B will each likely be approximately thermal states? (and, in particular, in the ‘total-pure state approach’, the total state will likely be approximately bithermal).

(What is meant here by ‘comparable size’ has, of course, to be encoded into the functional form of the densities of states $\sigma_S(\epsilon)$ and $\sigma_B(\epsilon)$. How this is done will be clear from the specific examples we discuss.)

We hope the answers we obtain below may be of interest in their own right and that the formalism we deploy to answer them may find a variety of other applications. But the immediate application we have in mind is to the theory of quantum black holes. In Section IX and in our two companion papers, [22, 23], we shall argue that our answers help to strengthen the case for, and give concrete form to, our matter-gravity entanglement hypothesis and particularly our entanglement picture of black hole equilibrium discussed in Section IB.

D. Answers

The key to answering our specific question, in the ‘traditional total microcanonical state’ approach is the formula (11) which we already gave above for the reduced density operator, ρ_S^{microc} , on S.

We claim that the appropriate replacement for this formula in the ‘modern total-pure state approach’ is

$$\rho_S^{\text{modapprox}} = M^{-1} \left(\sum_{\epsilon=\Delta}^{E_c} n_B(E-\epsilon) \sum_{i=1}^{n_S(\epsilon)} |\epsilon, i\rangle \langle \epsilon, i| + \sum_{\epsilon=E_c+\Delta}^E n_S(\epsilon) \sum_{i=1}^{n_B(E-\epsilon)} |\widetilde{\epsilon}, i\rangle \langle \widetilde{\epsilon}, i| \right). \quad (15)$$

On the right hand side of this equation, we continue to assume the spectrum to be slightly distorted in the way we explained before equation (6), n_S and n_B to be defined as in (4) and (5) and the sums to be over integral multiples of Δ , and we also continue to assume, as will be the case in our examples in Part 1, that σ_S and σ_B are monotonically increasing functions – defining E_c to be the energy value at which $\sigma_S(E_c) = \sigma_B(E - E_c)$. When $\epsilon > E_c$, the $|\widetilde{\epsilon}, i\rangle$ then denote the elements of an orthonormal basis of an $n_B(E - \epsilon)$ -dimensional subspace of the $(n_S(\epsilon))$ -dimensional energy- ϵ subspace of \mathcal{H}_S which will depend on Ψ . As we shall see, this dependence on Ψ will not matter for the developments in Part 1. We will postpone a full explanation of the way in which the subspace depends on Ψ to Section XII in Part 2.

It is important to notice that, as is easy to check, the constant, M , by which one needs to divide in order to normalize (15) has the same value, given by (7) and (9) (and as explained after those equations, equal to the total number of states of the totem with energy in the interval $[E, E + \Delta]$) as the constant, M , by which one needs to divide in order to normalize (11). Moreover, while the states, ρ_S^{microc} and $\rho_S^{\text{modapprox}}$, are clearly (usually) very different, both states share the same energy probability density, $P_S(\epsilon)$ (10). (There is of course a similar pair of equations to (11) and (15) with obvious reversals of the letters ‘S’ and ‘B’ and, in the case of (15), with E_c replaced by $E - E_c$.)

We now claim that the sense in which (15) is the appropriate replacement for (11) in the modern approach is then made clear by the following proposition, our argument for the correctness of which is given in (and is the main purpose of) Part 2:

Proposition. [21] *For a given, randomly chosen, pure state, Ψ , on the Hilbert space of our totem, with energy restricted to be in the range $[E, E + \Delta]$, the reduced density operator, ρ_S^{modern} of the system may, as far as physical*

quantities of interest are concerned, with very high probability, be considered to be very close to the $\rho_S^{\text{modapprox}}$ of (15) for the appropriate (i.e. to the chosen vector Ψ) $n_B(E - \epsilon)$ -dimensional subspaces of \mathcal{H}_S spanned by the $|\widetilde{\epsilon}, i\rangle$ (see above and Part 2). (And a similar statement of course holds with system, S, replaced by bath, B.)

What makes this proposition particularly useful is the fact that, while the $n_B(E - \epsilon)$ -dimensional subspaces (spanned by the $|\widetilde{\epsilon}, i\rangle$) of \mathcal{H}_S will depend on the choice of Ψ (in a way which we shall explain in Section XII in Part 2 where we point out, by the way, that they might themselves be said to be ‘random subspaces’) as is easy to see and as we shall illustrate in Part 1, the values of physical quantities of interest, such as the mean energy and the von Neumann entropy of the system S (see (16) below and Section IE) calculated using $\rho_S^{\text{modapprox}}$, do not depend on which $n_B(E - \epsilon)$ -dimensional subspaces they are. Therefore we can conclude that, to the extent that the approximation of ρ_S^{modern} by $\rho_S^{\text{modapprox}}$ is good (and we shall argue in Part 2 that, in our situations of interest, and when it is used for the purpose of calculating mean energy and entropy, it is very good) the actual values of these quantities must (with a very high probability) be largely independent of the choice of Ψ ! (Aside from mean energy, in fact we expect the entire energy probability density function, $P_S(\epsilon)$, will most likely be close to that of $\rho_S^{\text{modapprox}}$ and hence also, similarly, for higher moments of the energy.)

Above, we recall that, for an arbitrary density operator, ρ , the von Neumann entropy is given by the formula

$$S(\rho) = -\text{tr}(\rho \log \rho). \quad (16)$$

We remark that it is easy to see from a comparison between (11) and (15) that the above proposition implies the ‘Canonical Typicality’ result [9] of Goldstein, Lebowitz et al., thus fulfilling our promise in Section I to re-obtain the latter. For, in the relevant limit (see after Equation (12)) E_c in (15) will tend to E and therefore $\rho_S^{\text{modapprox}}$ (15) will tend to ρ_S^{microc} (11) which, in turn, will tend, by the traditional argument we reviewed in Section I, to a Gibbs state (namely the $\rho_{S,\beta}^{\text{Gibbs}}$ of (13) for β equal to the limiting value of N_B/E).

To start now to address our specific question, we first observe that, whatever the densities of states, σ_S and σ_B (provided only they are monotonically increasing) as long as the total energy, E , of our totem is finite, then, of course neither of the density operators, ρ_S^{microc} (11) and $\rho_S^{\text{modapprox}}$ (15), can be exactly thermal. To see this easily, it suffices to notice that the energy probability density, $P_S(\epsilon)$ (10), which these states share will obviously be zero for $\epsilon > E$, whereas, when $\sigma_S(\epsilon)$ is sufficiently slowly growing for $\rho_{S,\beta}^{\text{Gibbs}}$ (see (13)) to exist, the energy probability density for the Gibbs state (13) will obviously take the form

$$P_{S,\beta}^{\text{Gibbs}}(\epsilon) = Z_{S,\beta}^{-1} \sigma_S(\epsilon) \exp(-\beta\epsilon), \quad (17)$$

where $Z_{S,\beta}$ is as in (14), which (for a rising density of states σ_S) will be non-zero for all ϵ . However, one can ask whether ρ_S^{microc} and/or $\rho_S^{\text{modapprox}}$ can be *approximately thermal*, say at sufficiently low energies.

We shall find that, for physically ordinary densities of states such as (cf. the discussion around (12))

$$\sigma_S(\epsilon) = A_S \epsilon^{N_S}, \quad \sigma_B(\epsilon) = A_B \epsilon^{N_B}, \quad (18)$$

then, when system, S, and energy bath, B, are large and of comparable size – i.e. when N_S and N_B are both large, but comparably sized numbers – then neither ρ_S^{microc} nor $\rho_S^{\text{modapprox}}$ can even be approximately thermal. In particular, this is the case when both system and energy-bath densities of states are identical (i.e. when $A_S = A_B$ and $N_S = N_B$). Rather, we will show that, when system and energy bath are of comparable size (or identical) in the sense just explained, the energy probability density of both S and B will, instead of having the behaviour one would expect of a thermal state, deviate from the most likely distribution of energies between S and B according to a Gaussian probability distribution with width of the order of E divided by the square root of N_S (equivalently N_B).

On the other hand, we shall show that in certain well-defined senses, ‘approximately thermal’ states are obtained for system, S, and energy bath, B, both on the traditional total microcanonical state approach and also on the modern total pure state approach if they both have identical densities of states which either rise exponentially with energy or rise as ‘quadratic exponentials’ – i.e. each as the exponential of a constant times the square of the energy – the notion of ‘approximately thermal’ depending both on the approach (i.e. the traditional total microcanonical state approach or the modern total pure state approach) and also on the behaviour of the densities of states (i.e. on whether they rise as the exponential of energy or of energy squared). See especially the notions of ‘ E -approximately thermal’ and ‘ E -approximately semi-thermal’ introduced in Section III for the case of an exponentially rising density of states. (The extent to which these results generalize to non-identical densities of states is briefly discussed for the exponential case in Endnote [29] to Section III.)

E. Results on the origin of entropy

Although it is not indicated in our title, besides our main question concerning the origin of thermality, we shall be greatly concerned throughout the paper, with the origin of entropy. And we are particularly interested in understanding how the very large entropies of black holes come about.

To this end, we will obtain formulae (Equations (54), (55) in Section V and Equations (69) and (70) in Section VI) for the entropy of our system, S , on both traditional and modern approaches, when system and energy bath both have either identical exponential or identical quadratic exponential densities of states. (We will also obtain formulae for the mean energy of S and B .) In the traditional approach, this is simply the mean entropy of the reduced density operator of the system when the totem is in a microcanonical state with given energy, E . In the modern approach, we remark, first, that, for every pure totem state, whether or not S and B have identical densities of states, the system entropy is necessarily always equal to the energy-bath entropy and both of these quantities are, in fact, identical [48] with the {system}-{energy bath} entanglement entropy. Second, the value of the entropy in the modern case is to be interpreted, in the light of our proposition, as the value that the system entropy (= energy-bath entropy = {system}-{energy bath} entanglement entropy) of a randomly chosen totem pure state will, with very high probability, be very close to. One of the most significant of our overall conclusions, dependent on our proposition, which we argue for in Part 2, is the fact that there is such a value at all – i.e. the fact that, with our basic general assumptions and for system and energy-bath densities of states of the sorts we discuss, the vast majority of totem states will have a system entropy close to one single value, namely $-\text{tr}(\rho_S^{\text{modapprox}} \log(\rho_S^{\text{modapprox}}))$. In terms of the language of Quantum Information Theory, this may be stated in the following way (below we temporarily suspend our terminological conventions, calling both S and B ‘systems’ and our ‘totem’ the ‘total system’):

Given two comparably-sized large systems, (S and B), which are either uncoupled or weakly coupled, then (for physically reasonable densities of states and even some maybe physically unreasonable ones) if their total state is a random pure state, their degree of entanglement (as measured by their entanglement entropy, S) will, with high probability, be close to the single value $-\text{tr}(\rho_S^{\text{modapprox}} \log(\rho_S^{\text{modapprox}}))$.

(Similarly, we expect that the mean value of the energy of the system, S , will, with high probability, be close to the single value $-\text{tr}(\rho_S^{\text{modapprox}} H_S)$. Indeed we expect the full energy probability density function, $P_S(\epsilon)$ [and hence also other moments of the energy], of S to, be, with high probability, close to that of $\rho_S^{\text{modapprox}}$ [and similarly with S replaced by B].)

Our results are that, for a totem with total energy E , for identical exponentially rising densities of states, $\sigma_S(\epsilon) = \sigma_B(\epsilon) = ce^{b\epsilon}$, on the traditional approach, the entropy, S_S^{microc} , will be $bE/2$ (up to a logarithmic correction) while, on the modern approach, the entropy (i.e. the single value as discussed in the previous paragraph) $S_S^{\text{modapprox}}$, will be $bE/4$ (up to a logarithmic correction). For identical quadratic exponential densities of states, $\sigma_S(\epsilon) = \sigma_B(\epsilon) = Ke^{q\epsilon^2}$, we find that $S_S^{\text{microc}} = qE^2/2$ (up to a correction of order 1 in E), while $S_S^{\text{modapprox}}$ will be tiny (i.e. a term of order 1 in E). (In both traditional and modern cases and with both equal exponential and equal quadratic exponential densities of states the mean energy of both system and energy bath will, of course be $E/2$ – in the modern case, ‘mean energy’ here meaning the value, $-\text{tr}(\rho_S^{\text{modapprox}} H_S)$, that the mean energy of a random pure totem state will most likely be very close to.)

F. Outline of the rest of the paper

We shall give full details of the results outlined in Section ID in Part 1, the main sections of which comprise Section II, which discusses the case where the density of states of both system and energy bath goes as a power of the energy, Sections III and V, which discuss the exponential case, and Section VI, which discusses the quadratic exponential case. Section IV develops the mathematical formalism to enable efficient computation of the expected energy and entropy of system, S , and energy bath, B , for the states ρ_S^{microc} and $\rho_S^{\text{modapprox}}$ and this formalism is applied in Sections V and VI to obtain formulae for these quantities in the cases of exponential and quadratic exponential densities of states.

Two further sections, VII and VIII, discuss some further related matters and can be skipped on a first reading. Section VII discusses the special features of the entropy, in both modern and microcanonical cases, when the densities of states of system and energy bath are such that the energy probability density (10) is sharply peaked (as is, for example, the case for our power law densities of states) and derives some general formulae which enable us, e.g. to calculate the entropy for the states considered in Section II. In passing, we clarify the relation with some traditional work on the microcanonical ensemble (where peaks are normally presupposed) and dispel some myths. We also discuss the connection between the sum of the entropies of the partial states of system and energy bath with the totem entropy $\log(M)$. In Section VIII we point out that if some of our basic assumptions are relaxed, then the prospects for systems

to become hot become much less constrained and, in particular, there are ways in which a system can be hot while the totem is in a pure state which differ from the ‘modern’ scenarios we discuss below. In particular, we discuss the notion of ‘purification’ (closely related to ‘thermofield dynamics’).

The entropy formulae we obtain in Sections III, V and VI (as outlined at the end of Section IE) will play an important role in Section IX and in two companion papers [22, 23] where we discuss the application of the ideas and formulae of these sections to the theory of quantum black holes. In Section IX A, we point out an intriguing resemblance between our entropy and temperature formulae for quadratic exponential densities of states in the microcanonical strand of Section VI with Hawking’s energy and temperature formulae for (Schwarzschild) quantum black holes and point out an apparent lack of success for the modern strand of Section VI in modelling black holes. However, we argue that it is difficult to conclude anything decisive from these observations since (at least in a description in terms of a quantized Einsteinian metric) black holes presumably do not satisfy the basic assumptions underlying our results here – in particular our assumption (see Equation (2)) of weak coupling.

What seems more promising is a connection between the formulae and results for entropy and temperature which we obtain in Sections III and V for exponentially growing densities of states and scenarios in which quantum black holes are viewed as strong string-coupling limits of certain states of weakly coupled strings. In Section IX B and in our two companion papers, [22] and [23], we recall some of the existing work [38–41] in this direction, and point out that, despite its great computational success, what is computed in this work is the *degeneracy* of certain black hole states; the fact that the resulting degeneracy formulae happen to agree with the previously known values of black hole entropy does not seem to have been explained hitherto. We then go on to propose a modification of the existing string theory scenario, and in particular of the work of Susskind [38] and Horowitz and Polchinski [40, 41] based on the modern strand of the present paper and on our matter-gravity entanglement hypothesis and our entanglement picture of black hole equilibrium (see Section IB). We argue that this modified scenario, which is based on an understanding of black hole equilibrium states as strong string-coupling limits of equilibria involving a long string coupled to a stringy atmosphere, does offer an explanation of black hole entropy and thereby also a satisfactory resolution to the Information Loss Puzzle. The companion paper [22] gives a brief announcement of the main results of the present paper with a focus on the main results and formalism, as well as discussing further our matter-gravity entanglement hypothesis and outlining the application of that, with the results of Sections III and V, to this string scenario. The further companion paper [23] develops the string scenario further.

Part 2, which comprises Sections X, XI, XII and XIII, clarifies the meaning of Equation (15) and presents our arguments in favour of our proposition in Section ID. A fuller description of the contents of Part 2 is given towards the end of Section X.

Part 1: Results for power law, exponential and quadratic exponential (equal) densities of states

II. POWER-LAW DENSITIES OF STATES

If S and B have densities of states as in (18) then, by (9) and the remarks in the subsequent paragraph, we have that M , i.e. the total number of totem states with energy in $[E, E + \Delta]$, is given by

$$M = A_S A_B \Delta \int_0^E \epsilon^{N_S} (E - \epsilon)^{N_B} d\epsilon \quad (19)$$

which can be rewritten

$$M = A_S A_B \Delta E^{N_S + N_B + 1} B(N_S + 1, N_B + 1) \quad (20)$$

where $B(x, y)$ is the usual beta function (see e.g. [30]) – related to the gamma and factorial functions by

$$B(x + 1, y + 1) = \frac{\Gamma(x + 1)\Gamma(y + 1)}{\Gamma(x + y + 2)} = \frac{x!y!}{(x + y)!(x + y + 1)}. \quad (21)$$

(For fractional arguments, we take $x!$ to mean $\Gamma(x + 1)$.) On the other hand, the number of such totem states with system energy in an interval $[\epsilon, \epsilon + \delta]$ will, for suitable δ , be well-approximated by

$$P_S(\epsilon)\delta = A_S A_B \delta \Delta M^{-1} \epsilon^{N_S} (E - \epsilon)^{N_B}$$

$$= A_S A_B \delta \Delta M^{-1} E^{N_S + N_B} \left(\frac{\epsilon}{E} \right)^{N_S} \left(1 - \frac{\epsilon}{E} \right)^{N_B}. \quad (22)$$

Thus, combining (20), (22) and (21) we have that

$$P_S(\epsilon) = \frac{N_S + N_B + 1}{E} b(N_S; N_S + N_B, \frac{\epsilon}{E}) \quad (23)$$

where (see e.g. [26])

$$b(k; n, p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \quad (24)$$

is, when n and k are integers, the binomial distribution function which has the famous interpretation as the probability that n ‘Bernoulli’ trials, each with probability p for success and $q = 1 - p$ for failure, result in k successes and $n - k$ failures. In order to take advantage of the insight afforded by this connection with probability theory we shall (with negligible error when N_S and N_B are large) assume from now on that N_S and N_B , if not already integers, are replaced by their nearest integers.

First we notice that we may use the well-known connection between the binomial and the Poisson distribution to give an alternative derivation of the fact that, in the limit as E and N_B grow while N_S remains constant and the ratio N_B/E converges to β , S’s energy probability density, $P_S(\epsilon)$ (23), converges to the Gibbs energy probability density $P_{S,\beta}^{\text{Gibbs}}(\epsilon)$ (see (17) and (14)) with inverse temperature $\beta = N_B/E$ for $\sigma_S(\epsilon)$ as in (18) – the latter Gibbs energy probability density being given explicitly by

$$P_{S,\beta}^{\text{Gibbs}}(\epsilon) = \frac{\beta^{N_S+1} \epsilon^{N_S} e^{-\beta\epsilon}}{N_S!} \quad (25)$$

as one sees from (17) and (18) after easily checking from (14) and (18) that

$$Z_{S,\beta} = A_S^{-1} \frac{N_S!}{\beta^{N_S+1}}.$$

This convergence result is of course a special case (i.e. the case where S, as well as B, has a power-law density of states) of an easy corollary both of the traditional thermality result (on the total microcanonical state approach) and (bearing in mind the equality of the energy probability density for both (11) and (15)) of the ‘Canonical Typicality’ result of Goldstein Lebowitz et al. (on the ‘modern’ total pure state approach) which, as we discussed in Section I, both hold in the same limit; we shall see shortly that the alternative proof which we next give for this corollary easily implies an alternative proof to the traditional thermality result itself and thus also, by a remark we made in Section ID to an alternative argument for ‘Canonical Typicality’ when the system and energy-bath densities of states both have power-law form.

As Feller puts it in [26], “If n is large and p is small, whereas the product $\lambda = np$ is of moderate magnitude” then the binomial distribution goes over to the Poisson distribution, i.e.

$$b(k; n, p) \simeq \frac{\lambda^k}{k!} e^{-\lambda}. \quad (26)$$

In particular (cf. e.g. [27]) for fixed k , the right hand side of (26) is the limit of $b(n; k, p)$ as $n \rightarrow \infty$ while $p \rightarrow 0$ in such a way that $np \rightarrow \lambda$. From this, and (23), we easily conclude that the limit, as $E \rightarrow \infty$ while $N_B/E \rightarrow \beta$ with N_S fixed, of $P_S(\epsilon)$ is equal to $P_{S,\beta}^{\text{Gibbs}}(\epsilon)$ (25). So, to summarize, in the appropriate limit of a large energy bath, the energy probability density of S goes over to the energy probability density of the appropriate Gibbs state;

$$P_S(\epsilon) \rightarrow P_{S,\beta}^{\text{Gibbs}}(\epsilon) = \frac{\beta^{N_S+1} \epsilon^{N_S} e^{-\beta\epsilon}}{N_S!}. \quad (27)$$

We remark that, by inspecting (17) and (14) this is easily seen to be equivalent to the statement that, in the same limit,

$$M^{-1} n_B(E - \epsilon) \rightarrow Z_{S,\beta}^{-1} e^{-\beta\epsilon}$$

and, by inspecting (11) and (13), one easily sees that this implies that in the same limit

$$\rho_S^{\text{microc}} \rightarrow \rho_{S,\beta}^{\text{Gibbs}}$$

thus providing the alternative proof, which we promised in Section I, of the traditional result on the thermality of a small system in contact with an energy bath in the traditional limit of a large energy bath, in the case where both the energy bath, B, and system, S, have a power-law density of states (and by our remark in Section ID thus also providing an alternative proof of ‘Canonical Typicality’ for such S and B).

We will now demonstrate, however, that, when S and B are of *comparable size*, then, if they both have power-law densities of states as in (18), both the total microcanonical state approach and the ‘modern’ approach (i.e. with a total pure state) predict that the reduced density operators of each of S and B will be quite different from thermal! We shall show this by showing that the energy probability density of each of S and B (which we again recall from the paragraph after (15) is the same in each approach) will have a quite different form from the thermal form of $P_{S,\beta}^{\text{Gibbs}}(\epsilon)$.

First we notice that, when k is a fixed fraction, pn , of n (in such a way that $0 < p < 1$ and also pn is an integer) then, if p and n are regarded as fixed, the binomial distribution function (24) $b(pn; n, p')$ is maximized when $p' = p$ and we easily obtain the approximation [28] (now writing $p' = p + x$)

$$b(pn; n, p + x) \simeq \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{nx^2}{2p(1-p)}\right). \quad (28)$$

(28) is obtained by expressing the left hand side in terms of factorials and powers according to (24). We then adopt Stirling’s approximation, $N! \simeq \sqrt{2\pi} N^{N+\frac{1}{2}} e^{-N}$ for each of the factorials and, introducing $q = 1 - p$, write the term $(p+x)^{np}(1-p-x)^{n(1-p)}$ as $p^{np}q^{nq}$ times $(1+x/p)^{np}(1-x/q)^{nq}$ and approximate the latter by $\exp(-nx^2/2pq)$. Clearly, as long as n is extremely large and p is not extremely close to zero or 1, then this will be an excellent approximation.

Combining (23) with the definition of p before (28) we see that, if we identify n with $N_S + N_B$, then

$$P_S(\epsilon) = \frac{n+1}{E} b\left(pn; n, \frac{\epsilon}{E}\right)$$

where

$$p = \frac{N_S}{N_S + N_B} \quad (29)$$

and that, provided n is extremely large and S and B are of ‘comparable size’, which of course, in view of (29), corresponds exactly to p not being extremely close to zero or 1, then, by (28), to a high degree of accuracy, we will have the approximation

$$P_S(\epsilon) \simeq \frac{1}{E} \sqrt{\frac{\gamma}{\pi}} \exp\left(-\gamma \frac{(\epsilon - \epsilon_0)^2}{E^2}\right) \quad (30)$$

i.e. a Gaussian with a peak located (See Section VII A for an alternative perspective on Equation (31)) at

$$\epsilon_0 = pE \quad (p \text{ as in (29)}) = \frac{N_S}{N_S + N_B} E \quad (31)$$

and

$$\gamma = \frac{n}{2p(1-p)} = \frac{(N_S + N_B)^3}{2N_S N_B} \quad (32)$$

and there will of course be an obvious counterpart formula for the energy probability density, P_B , of B, similar to the above formula but with p replaced by $1 - p$. (This of course changes the value of ϵ_0 but not of γ .) So the energy of S will be in a Gaussian band around a most likely energy of ϵ_0 , the energy of B will be in a Gaussian band around a most likely energy of $E - \epsilon_0$, each having the same width which will be E divided by a number (i.e. $\sqrt{2\gamma}$) which is of the order of the square root of either of the (comparable!) numbers N_S, N_B . Moreover it is easy to see that, in both the traditional microcanonical and the modern total pure state approaches, the two energy probability densities will be perfectly anticorrelated – i.e. when S has energy in a small interval around ϵ , then B will have energy in a similar small interval around $E - \epsilon$.

Above, by ‘width’ we mean the standard deviation, \mathfrak{s} , from the mean of the energy probability density, i.e.

$$\mathfrak{s} = (\overline{\epsilon^2} - \bar{\epsilon}^2)^{1/2}$$

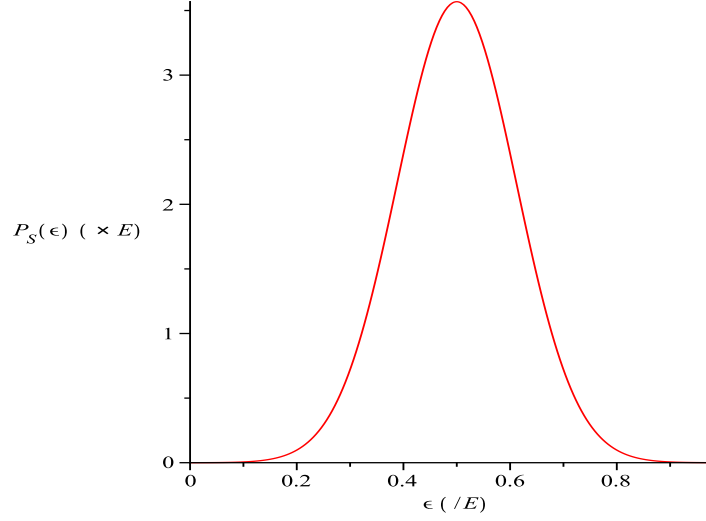


FIG. 1: Plot of the energy probability density (23), $P_S(\epsilon)$, in the case S and B have the same density of states $\sigma(\epsilon) = A\epsilon^N$ for the (‘unusually’ small) value $N = 10$

where

$$\overline{\epsilon^n} = \int_0^E \epsilon^n P_S(\epsilon) d\epsilon \quad (33)$$

($= \text{tr}(\rho_S^{\text{microc}} H_S^n) = \text{tr}(\rho_S^{\text{modapprox}} H_S^n)$ – cf. Section IV).

(In the special case where $N_S = N_B = N$ say, one sees that γ becomes $4N$ so the width, \mathfrak{s} , will be $E/2\sqrt{2N}$.) As we anticipated, this is a qualitatively very different behaviour from the energy probability density of thermal states and we conclude therefore, as promised, that, in both the traditional total microcanonical and the modern total pure state approaches, the reduced density operators of S and B must, when, S and B are of comparable size, be quite different in character from thermal density operators. To illustrate this point, we include a figure (Figure 1) for the energy probability density, $P_S(\epsilon)$, in the case S and B have the same density of states $\sigma(\epsilon) = A\epsilon^N$ for the (unrealistically small) value $N = 10$ and a comparison figure, Figure 2, showing the the energy probability density, $P_{S,\beta}^{\text{Gibbs}}(\epsilon)$ for a thermal state at the inverse temperature, $\beta = 2(N+1)/E$, chosen so that the mean energy takes the same value, $E/2$ – again in the case $N = 10$.

For the sake of a quantitative result, we note that, for general N , the width, \mathfrak{s} , of the energy probability density, $P_{S,\beta}^{\text{Gibbs}}(\epsilon)$, of this comparison thermal state is (as is easily calculated) $E/2\sqrt{N+1}$ – i.e. (to a very good approximation for large N) a factor of $\sqrt{2}$ wider than the width of $P_S(\epsilon)$ while the height is (again by an easy calculation) a factor of $\sqrt{2}$ smaller.

We shall postpone to Section VII a calculation of the (microcanonical and modern) entropies of S and B for general N_S and N_B . Suffice it to remark that, like the width, \mathfrak{s} , the microcanonical entropy of S, differs, in general, from its value in the comparison thermal state at inverse temperature $\beta = 2(N+1)/E$, albeit the difference is just a ‘small’ constant (it is smaller by $\simeq \log 2/2$) independent of ϵ .

Finally, we remark that, in this power-law density-of-states case, it is clear from the developments in this section that the ‘canonical’ (i.e. thermal) behaviour of ρ_S^{microc} (or indeed of $\rho_S^{\text{modapprox}}$) in the case that the system, S, is very much smaller than the energy bath, B, may be reconciled with the above-discussed Gaussian behaviour, when S and B are of comparable size, in that the relationship between the two may be regarded as an instance of the well-

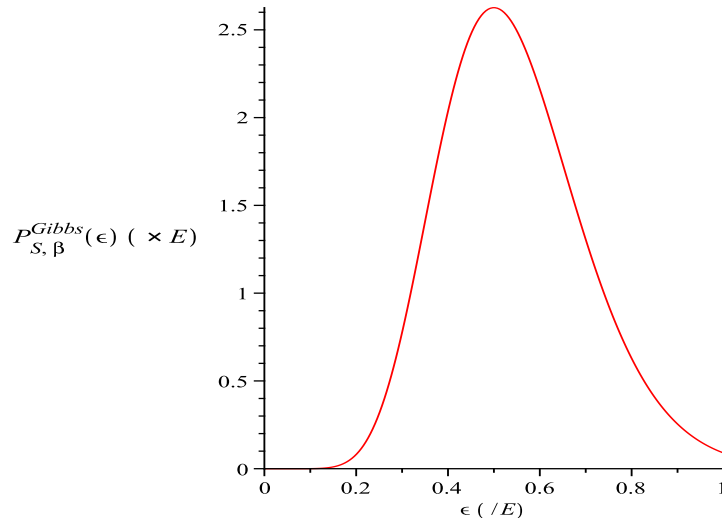


FIG. 2: Plot of the energy probability density, $P_{S, \beta}^{Gibbs}(\epsilon)$ for the thermal state at inverse temperature, β , on our system, S, with density of states $\sigma(\epsilon) = A\epsilon^N$, for the same (‘unusually’ small) value $N = 10$ and for $\beta = 22/E$ (i.e. the value of β for which the mean energy is $E/2$). To be contrasted with the $P_S(\epsilon)$ of Figure 1.

known relationship (see e.g. [26] or [27]) between the Poisson and Gaussian distributions in probability theory. (This obviously easily follows from the way we derived, above, both the canonical behaviour and the Gaussian behaviour as limits of the binomial distribution.)

III. EXPONENTIALLY RISING DENSITIES OF STATES

We now turn to discuss the quite different behaviour of the reduced density operators ρ_S^{microc} and $\rho_S^{\text{modapprox}}$ when the densities of states of S and B increase exponentially. We shall confine our interest here to the case where both densities of states, σ_S and σ_B , behave as $ce^{b\epsilon}$ with the same constants c and b in each expression:

$$\sigma_S(\epsilon) = ce^{b\epsilon}, \quad \sigma_B(\epsilon) = ce^{b\epsilon}. \quad (34)$$

We remark, however, that, as may quite easily be checked, allowing different values of c (say c_S in the first formula and c_B in the second) will not essentially change our conclusions [29].

The normalization constant M is now easily seen – either by using (9) or, on recalling (4), by using (7) – to be given by

$$M = c^2 e^{bE} E \Delta. \quad (35)$$

We note that this will be large provided neither c nor Δ are ‘too small’ and provided also

$$bE \gg 1 \quad (36)$$

which will hold in cases of interest.

The formula, (11) for ρ_S^{microc} is then easily seen to coincide with the formula, (13), for a thermal density operator $\rho_{S, \beta}^{Gibbs}$, for the density of states $\sigma_S(\epsilon)$ as in (34) at inverse temperature $\beta = b$, provided the latter formula is modified

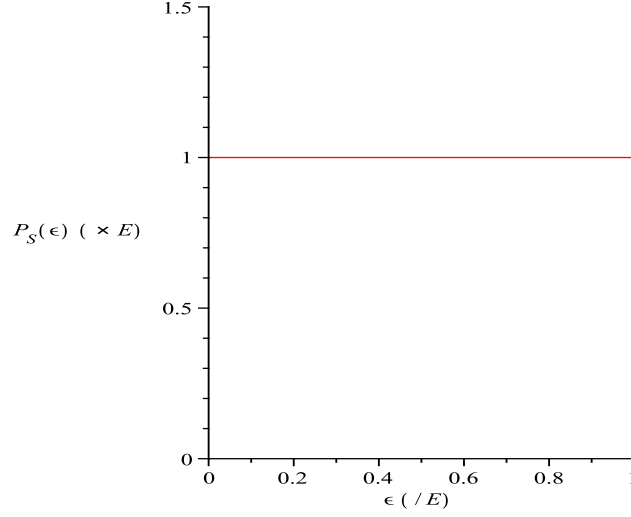


FIG. 3: Plot of the energy probability density (37), $P_S(\epsilon)$, in the case S and B have the same density of states $\sigma(\epsilon) = ce^{b\epsilon}$

so that the sum over ϵ is truncated at the upper energy, E and the partition function, $Z_{S,\beta}$, is replaced by cE . Of course, the un-truncated formula (13) will only make mathematical sense for $\beta > b$. Nevertheless, the reduced density operator ρ_S^{microc} (and similarly also ρ_B^{microc}) clearly deserves to be called an approximately thermal state at inverse temperature b . (This will generalize from equal systems to comparably sized systems if, by this, we mean systems with densities of states with unequal c_S and c_B as discussed in Endnote [29]). We shall refer to the relevant notion of being approximately thermal here as being *E-approximately thermal*.

Turning from the traditional total microcanonical state approach to the modern total pure state approach, we see, on substituting (34) into (15) and noting that E_c will obviously become $E/2$, that the ϵ -summand in (15) still agrees with the ϵ -summand in (13) at inverse temperature $\beta = b$ up to energy $E/2$ and, moreover, as always (cf. after Equation (15)) the system energy probability density of $\rho_S^{\text{modapprox}}$ is equal to that of ρ_S^{microc} and thus it agrees with the energy probability density of a Gibbs state, for the same density of states, up to energy E . We shall refer to the relevant notion of being approximately thermal here (i.e. agreement of the summand in the formula (15) with the summand in the formula (13) up to $\epsilon = E/2$ – with a suitable change in the value of $Z_{S,\beta}$ – and agreement of the energy probability density up to E) as being *E-approximately semi-thermal*.

We note here that, with the densities of states as in (34), the energy probability density $P_S(\epsilon)$, which we recall by (10) is given in general by

$$P_S(\epsilon) = \frac{\Delta}{M} \sigma_S(\epsilon) \sigma_B(E - \epsilon),$$

will, with M as in (35) and σ_S and σ_B as in (34), reduce to

$$P_S(\epsilon) = \frac{1}{E}. \quad (37)$$

See Figure 3. Of course (cf. the paragraph after Equation (32)) the energies of S and B will, again, be perfectly anticorrelated.

Similar results to the the above results for S will obviously hold for B. We thus conclude, in fulfillment of our promise (cf. the start of Section IC) that, with the appropriate meaning in each case for the expression “approximately thermal”, as above, when the densities of states of S and B take the exponential form of (34) then – in contrast to

the situation for power-law densities of states discussed in Section II – on both the traditional total microcanonical state approach and also on the modern total pure state approach, the reduced density operators of both S and B will be approximately thermal (at inverse temperature $\beta = b$) in an appropriate sense.

IV. GENERAL FORMULAE FOR MEAN ENERGY AND FOR ENTROPY

It is interesting (especially in preparation for the discussion in Section IX and in our companion papers [22, 23] about the connection with quantum black hole physics both of the results in Sections III and V concerning densities of states which grow exponentially with energy and, in Section VI, concerning those which grow as quadratic exponentials in the energy) to calculate the mean energy, $\bar{\epsilon}$, and also the von Neumann entropy, S , for each of the density operators, $\rho_S^{\text{microc}}, \rho_S^{\text{modapprox}}$ (and also for $\rho_B^{\text{microc}}, \rho_B^{\text{modapprox}}$). The former density operator will just give the usual mean energy and von Neumann entropy (defined as in (16)) of our system, S, when the totem is in the microcanonical ensemble. The latter density operator will, according to our proposition in Section ID, give an energy value and an entropy value which will very probably be very close to the mean value of the energy and the entropy of our system, S, when the totem is in a random pure state.

By (11), (15), and (7), we have, in general, that, with obvious notation,

$$\bar{\epsilon}_S^{\text{microc}} := \text{tr}(\rho_S^{\text{microc}} H_S) = M^{-1} \sum_{\epsilon=\Delta}^E \epsilon n_S(\epsilon) n_B(E - \epsilon).$$

Similarly

$$\bar{\epsilon}_B^{\text{microc}} := \text{tr}(\rho_B^{\text{microc}} H_B) = M^{-1} \sum_{\epsilon=\Delta}^E \epsilon n_B(\epsilon) n_S(E - \epsilon)$$

and one easily sees that necessarily, $\bar{\epsilon}_B^{\text{microc}} = E - \bar{\epsilon}_S^{\text{microc}}$. Moreover, we have

$$\bar{\epsilon}_S^{\text{modapprox}} = \text{tr}(\rho_S^{\text{modapprox}} H_S) \quad (38)$$

$$= M^{-1} \left(\sum_{\epsilon=\Delta}^{E_c} \epsilon n_B(E - \epsilon) n_S(\epsilon) + \sum_{\epsilon=E_c+\Delta}^E \epsilon n_S(\epsilon) n_B(E - \epsilon) \right), \quad (39)$$

which is easily seen to be equal to $\bar{\epsilon}_S^{\text{microc}}$. Similarly, $\bar{\epsilon}_B^{\text{modapprox}} = \bar{\epsilon}_B^{\text{microc}}$.

On the other hand, by (11) and (16), we will have

$$S_S^{\text{microc}} := -\text{tr}(\rho_S^{\text{microc}} \log \rho_S^{\text{microc}}) = -M^{-1} \sum_{\epsilon=\Delta}^E n_S(\epsilon) n_B(E - \epsilon) \log(M^{-1} n_B(E - \epsilon)) \quad (40)$$

and by (15) and (16)

$$\begin{aligned} S_S^{\text{modapprox}} &:= -\text{tr}(\rho_S^{\text{modapprox}} \log \rho_S^{\text{modapprox}}) \\ &= -M^{-1} \left(\sum_{\epsilon=\Delta}^{E_c} n_S(\epsilon) n_B(E - \epsilon) \log(M^{-1} n_B(E - \epsilon)) + \sum_{\epsilon=E_c+\Delta}^E n_S(\epsilon) n_B(E - \epsilon) \log(M^{-1} n_S(\epsilon)) \right) \end{aligned} \quad (41)$$

and similarly with S replaced by B. We remark that it is not difficult to see from (15) and the counterpart equation for $\rho_B^{\text{modapprox}}$ that $S_S^{\text{modapprox}}$ will necessarily equal $S_B^{\text{modapprox}}$. This is of course consistent with the fact that, by the general result recalled in Endnote [48], for any pure totem state, Ψ , we necessarily have that the von Neumann entropies of the resulting reduced density operators ρ_S^{modern} and ρ_B^{modern} will necessarily be equal. After all, as we claim in our Proposition in Section ID and argue in Part 2, for random Ψ , $\rho_S^{\text{modapprox}}$ most probably gives a very good approximation of ρ_S^{modern} and $\rho_B^{\text{modapprox}}$ of ρ_B^{modern} .

By referring to the second equality in (10), it is easy to see that the formulae for S_S^{microc} and $S_S^{\text{modapprox}}$ in (40) and (41) may be rearranged to give the following useful alternative expressions:

$$S_S^{\text{microc}} = \Delta \sum_{\epsilon=\Delta}^E P_S(\epsilon) \log \left(\frac{n_S(\epsilon)}{P_S(\epsilon)\Delta} \right), \quad (42)$$

$$S_S^{\text{modapprox}} = \Delta \left(\sum_{\epsilon=\Delta}^{E_c} P_S(\epsilon) \log \left(\frac{n_S(\epsilon)}{P_S(\epsilon)\Delta} \right) + \sum_{\epsilon=E_c+\Delta}^E P_S(\epsilon) \log \left(\frac{n_B(E-\epsilon)}{P_S(\epsilon)\Delta} \right) \right). \quad (43)$$

Referring to (4) and making the replacement (8) (and with the proviso made in the cautionary remark in [14]) we see that (42) and (43) have, as their continuum versions:

$$S_S^{\text{microc}} = \int_0^E P_S(\epsilon) \log \left(\frac{\sigma_S(\epsilon)}{P_S(\epsilon)} \right) d\epsilon, \quad (44)$$

$$S_S^{\text{modapprox}} = \int_0^{E_c} P_S(\epsilon) \log \left(\frac{\sigma_S(\epsilon)}{P_S(\epsilon)} \right) d\epsilon + \int_{E_c}^E P_S(\epsilon) \log \left(\frac{\sigma_B(E-\epsilon)}{P_S(\epsilon)} \right) d\epsilon. \quad (45)$$

We notice, in passing, that the absence of the quantity Δ (or of any quantity that scales with Δ) in the formulae S_S^{microc} and $S_S^{\text{modapprox}}$ shows us the interesting fact that (for Δ in an appropriate not-too-large and not-too-small range, and to what, in typical applications will be an extremely good approximation) neither of these entropies depends on Δ !

Finally, further useful insight concerning the form of Equations (42) and (43) can be had by noticing that they can alternatively be derived as corollaries of the following easily proved Lemma, which we will also need to refer to in Section XIII in Part 2.

Lemma: *Given density operators ρ_1, ρ_2, \dots on Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2, \dots$ respectively, with von Neumann entropies $S(\rho_1), S(\rho_2), \dots$ and given positive real numbers $\lambda_1, \lambda_2, \dots$ with $\sum_i \lambda_i = 1$. Then the density operator*

$$\rho = \lambda_1 \rho_1 \oplus \lambda_2 \rho_2 \oplus \dots$$

on the direct sum Hilbert space $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \dots$ will have an entropy, S , given by

$$S = \sum_i \lambda_i S(\rho_i) - \sum_i \lambda_i \log \lambda_i. \quad (46)$$

To apply this lemma to the calculation of S_S^{microc} and $S_S^{\text{modapprox}}$ (for general densities of states) we first notice that (11) and (15) can be rewritten as

$$\rho_S^{\text{micro}} = \oplus_{\epsilon=\Delta}^E \lambda_\epsilon \rho_\epsilon \quad (47)$$

and

$$\rho_S^{\text{modapprox}} = \oplus_{\epsilon=\Delta}^{E_c} \lambda_\epsilon \rho_\epsilon + \oplus_{\epsilon=E_c+\Delta}^E \tilde{\lambda}_\epsilon \tilde{\rho}_\epsilon \quad (48)$$

where

$$\rho_\epsilon = n_S(\epsilon)^{-1} \sum_{i=1}^{n_S(\epsilon)} |\epsilon, i\rangle \langle \epsilon, i| \quad (49)$$

and

$$\tilde{\rho}_\epsilon = n_B(E-\epsilon)^{-1} \sum_{i=1}^{n_B(E-\epsilon)} |\widetilde{\epsilon}, i\rangle \langle \widetilde{\epsilon}, i| \quad (50)$$

where, $|\epsilon, i\rangle$ and $|\widetilde{\epsilon}, i\rangle$ are as in (11) and (15), and where (recalling that the sums in (11) and (15) are over energies, ϵ , which are integral multiples of Δ)

$$\lambda_\epsilon = \tilde{\lambda}_\epsilon = P_S(\epsilon)\Delta \quad (51)$$

where $P_S(\epsilon)$ is the energy probability density (10).

We also easily see from (49) and (50) that, in general

$$S(\rho_\epsilon) = \log(n_S(\epsilon)) \quad \text{and} \quad S(\tilde{\rho}_\epsilon) = \log(n_B(E-\epsilon)). \quad (52)$$

Equations (42) and (43) now follow by simple applications of the formula (46) or of our above lemma to (47) and (48).

V. FORMULAE FOR MEAN ENERGY AND ENTROPY FOR EXPONENTIALLY RISING DENSITIES OF STATES

Specialising to n_S and n_B given by $n_S(\epsilon) = \sigma_S(\epsilon)\Delta$ and $n_B(\epsilon) = \sigma_B(\epsilon)\Delta$ with σ_S and σ_B as in (34), we have that

$$\bar{\epsilon}_S^{\text{microc}} = \bar{\epsilon}_S^{\text{modapprox}} = E/2, \quad (53)$$

and similarly with S replaced by B. These results for the mean energies of S and B are of course anyway obvious since the assumption of very weak coupling (see (2) and the subsequent paragraph) implies that the mean energies of S and B will add to E , while (34) is symmetric under the replacement of S by B. However, we remark that (53) turns out to remain exactly unchanged even when the densities of states are generalized so as to have different pre-factors c_S and c_B (see Endnote [29]). (Returning to the case of equal densities of states) we caution that the mean energies are just that, averages; they are not in any sense ‘most likely’ energies. In fact, as we saw in Section III, the energy probability density (see (37) and Figure 3) is flat!

It is also straightforward to calculate, using the formulae of Section IV, that the entropies take the values

$$S_S^{\text{microc}} = bE/2 + \log(cE), \quad (54)$$

$$S_S^{\text{modapprox}} = bE/4 + \log(cE), \quad (55)$$

and similarly with S replaced by B. (Again, see Endnote [29] for the generalization to different prefactors, c_S and c_B , in the first and second formulae of (34)).

In particular, inserting the formulae (34) and (37) for σ_S and P_S in (42) and (43) we obtain

$$\begin{aligned} S_S^{\text{microc}} &= \sum_{\epsilon=\Delta}^E \frac{\Delta}{E} (b\epsilon + \log(c\Delta) - \log(\Delta/E)) \\ &= \frac{\Delta}{E} \left(b\Delta \frac{(E/\Delta)(E/\Delta + 1)}{2} + (E/\Delta) \log(cE) \right) \\ &\simeq \frac{bE}{2} + \log(cE) \end{aligned} \quad (56)$$

while (assuming E/Δ is even)

$$\begin{aligned} S_S^{\text{modapprox}} &= \sum_{\epsilon=\Delta}^{E/2} \frac{\Delta}{E} (b\epsilon + \log(c\Delta) - \log(\Delta/E)) + \sum_{\epsilon=E/2+\Delta}^E \frac{\Delta}{E} (b(E-\epsilon) + \log(c\Delta) - \log(\Delta/E)) \\ &= 2 \sum_{\epsilon=\Delta}^{E/2} \frac{\Delta}{E} (b\epsilon + \log(c\Delta) - \log(\Delta/E)) \\ &= 2 \frac{\Delta}{E} \left(b\Delta \frac{(E/2\Delta)(E/2\Delta + 1)}{2} + (E/\Delta) \log(cE) \right) \\ &\simeq \frac{bE}{4} + \log(cE) \end{aligned} \quad (57)$$

which are the formulae (54) and (55). In the calculations above, we need to recall that the sums in (11) and (15), and hence also in the direct sums in (47) and (48) and in the above equations, are over ϵ values which are positive-integer multiples of Δ .

We remark that the leading behaviour of S_S^{microc} (54) (i.e. the term, $bE/2$, which remains when we ignore the logarithmic terms in (54)) arises, in (say) the continuum version, (44), of our general formula for S_S^{microc} by replacing the logarithm in this formula by its ‘main part’, by which we mean the exponent, $b\epsilon$, in the formula, (34) $\sigma_S(\epsilon) = ce^{b\epsilon}$. Similarly, the leading behaviour of $S_S^{\text{modapprox}}$ (i.e. the term $bE/4$ in (55)) arises by setting $E_c = E/2$ in (45) and noticing that the main parts of the two logarithms in this formula are (in order) $b\epsilon$ and $b(E/2 - \epsilon)$.

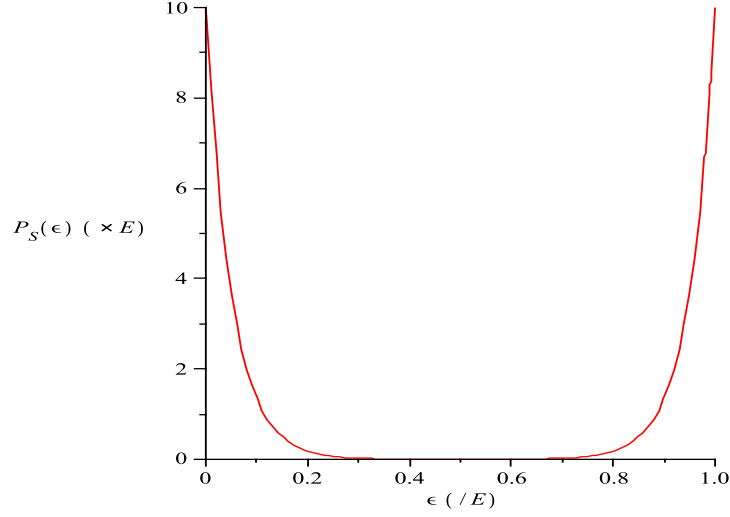


FIG. 4: Plot of the energy probability density (60), $P_S(\epsilon)$, in the case S and B have the same density of states $\sigma(\epsilon) = Ke^{q\epsilon^2}$ for the (unrealistically small) value $qE^2 = 10$.

VI. DENSITIES OF STATES WHICH GROW AS QUADRATIC EXPONENTIALS

Next we discuss the behaviour of the reduced density operators, ρ_S^{microc} and $\rho_B^{\text{modapprox}}$, when the densities of states of S and B each increase as the exponential of a constant times the square of the energy. We shall confine our interest to the case where both densities of states, σ_S and σ_B , behave as $Ke^{q\epsilon^2}$ with the same constants, K and q :

$$\sigma_S(\epsilon) = Ke^{q\epsilon^2}, \quad \sigma_B(\epsilon) = Ke^{q\epsilon^2} \quad (58)$$

and shall just discuss the cases of ρ_S^{microc} and $\rho_S^{\text{modapprox}}$ – those of ρ_B^{microc} and $\rho_B^{\text{modapprox}}$ obviously being similar.

We shall assume that

$$qE^2 \gg 1. \quad (59)$$

The energy probability density, $P_S(\epsilon)$ (10), now takes the form

$$P_S(\epsilon) = \frac{\Delta}{M} K^2 e^{qE^2} e^{-2q\epsilon(E-\epsilon)}. \quad (60)$$

and we sketch its graph in Figure 4.

We note that it is symmetric about $E/2$ and also, in view of (59), $P_S(\epsilon)$ is very close to zero except when ϵ is close to 0 or to E , where it is well approximated by the exponentially decaying function $\frac{\Delta}{M} K^2 e^{qE^2} e^{-2qE\epsilon}$ (near $\epsilon = 0$), and by the exponentially rising function $\frac{\Delta}{M} K^2 e^{qE^2} e^{2qE(\epsilon-E)}$ (near $\epsilon = E$). Approximating the integral from 0 to E of $P_S(\epsilon)$ by the sum of the (equal) integrals (from 0 to ∞ and from $-\infty$ to E) of these exponential approximations, and demanding that the result must equal 1 we see that $P_S(\epsilon)$ will be well approximated on its domain $[0, E]$ by

$$P_S(\epsilon) \simeq qE(e^{-2qE\epsilon} + e^{2qE(\epsilon-E)}) \quad (61)$$

from which we infer that the normalization constant, M , will be well approximated [42] by

$$M \simeq K^2 \frac{e^{qE^2}}{qE} \Delta. \quad (62)$$

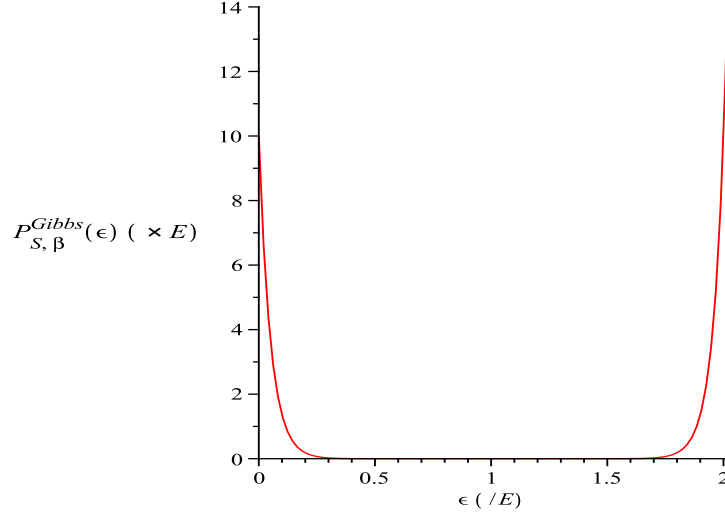


FIG. 5: Comparison figure for the (non-normalizable!) “energy probability density” (64) of the “thermal state”, $P_{S, \beta}^{Gibbs}(\epsilon)$ at inverse temperature $\beta = 2qE$, again for the value $qE^2 = 10$

We pause here to record that (cf. (53)) either by calculation or by the above-noted symmetry of $P_S(\epsilon)$, we will obviously have that the mean energy of both S and E will be $E/2$:

$$\bar{\epsilon}_S^{\text{microc}} = \bar{\epsilon}_S^{\text{modapprox}} = E/2, \quad (63)$$

However, (cf. our remark after Equation (53)) even more emphatically than in the case of exponentially rising densities of states, these are *not* ‘most likely energies’. In fact, the energy probability density, $P_S(\epsilon)$, tells us that the energy (say of S) will be highly likely, and with equal likelihoods, either to be close to 0 or to be close to E – and highly unlikely to be close to $E/2$ (and similarly for B). In addition of course (cf. after Equation (32) and Equation (37)) the energy of S and the energy of B will be perfectly anticorrelated. So when the energy of S is near 0, the energy of B will be near E , and when the energy of S is near E , the energy of B will be near 0.

In analogy to what we did in Sections III, V, we next wish to compare the formulae (60) and (61) for $P_S(\epsilon)$ with

$$P_{S, \beta}^{Gibbs}(\epsilon) = C e^{-\beta \epsilon} K e^{q \epsilon^2}, \quad (64)$$

where C is a suitable constant, which, but for the fact that it is not integrable on the interval $(0, \infty)$ (for any value of β !) would deserve to be called ‘the energy probability density of a thermal state at inverse temperature β ’ (cf. (17)) for the same density of states (58). If $\beta E \gg 1$, then, on the interval $[0, \beta/q]$, $P_{S, \beta}^{Gibbs}(\epsilon)$ will be very close to zero except when ϵ is close to 0 or to β/q , where it will be well approximated by the exponential decay $C e^{-\beta \epsilon}$ (near 0), and by the exponentially rising function $C e^{\beta(\epsilon - \beta/q)}$ (near $\epsilon = \beta/q$). Put otherwise, on the interval $[0, \beta/q]$, $P_{S, \beta}^{Gibbs}(\epsilon)$ will take the approximate form

$$P_{S, \beta}^{Gibbs}(\epsilon) \simeq C K (e^{-\beta \epsilon} + e^{\beta(\epsilon - \beta/q)}). \quad (65)$$

Beyond $\epsilon = \beta/q$, $P_{S, \beta}^{Gibbs}(\epsilon)$ will, of course, grow rapidly. If we now choose to make the identification

$$\beta = 2qE, \quad (66)$$

we see that (65) can be written

$$P_{S,\beta}^{\text{Gibbs}}(\epsilon) \simeq CK(e^{-\beta\epsilon} + e^{\beta(\epsilon-2E)}) \quad (67)$$

while (61) can be written

$$P_S(\epsilon) \simeq qE(e^{-\beta\epsilon} + e^{\beta(\epsilon-E)}) \quad (68)$$

and comparison of (67) and (68) or a glance at Figures 4 and 5, immediately shows that these resemble one-another closely – each having an equally-rapidly exponentially decaying peak located near $\epsilon = 0$ and each having an equally rapidly exponentially rising, equal-sized second peak – the only discrepancy being that, in the case of $P_S(\epsilon)$, the second peak is located near $\epsilon = E$ while, in the case of $P_{S,\beta}^{\text{Gibbs}}(\epsilon)$, it is located near $\epsilon = 2E$. Thus, at low energies – and indeed at all energies, ϵ , up to a little below E – the energy probability density, $P_S(\epsilon)$, is closely approximated by the thermal energy probability density for $\beta = 2qE$, while there is a qualitative resemblance between $P_S(\epsilon)$ on its full interval $[0, E]$ and $P_{S,\beta}^{\text{Gibbs}}(\epsilon)$ on the interval $[0, 2E]$ (with the above-mentioned quantitative discrepancy that the second peak in $P_S(\epsilon)$ occurs near E while the second peak in $P_{S,\beta}^{\text{Gibbs}}(\epsilon)$ occurs near $2E$).

Moreover, one may easily check that, except for discrepancies corresponding to the above discrepancy for the energy probability densities, the reduced density operators, ρ_S^{microc} (defined as in (11)) and $\rho_S^{\text{modapprox}}$ (defined as in (15)), respectively, will stand in relation to $\rho_{S,\beta}^{\text{Gibbs}}$ (defined as in (13)) for $\beta = 2qE$, in a similar way to the relationships which we termed ‘ E -approximately thermal’ and ‘ E -approximately semi-thermal’ in Section III.

In conclusion, except for the discrepancy pointed out above, we may say that, in contrast again to the situation for power-law densities of states and with many similarities (but also a few differences) to what we found in Sections III, V, for densities of states which grow exponentially with energy, also densities of states which grow, (58), as quadratic exponentials lead to reduced density operators on S and B which are, in the sense we have explained above, approximately thermal.

Next we turn to calculate the von Neumann entropies of ρ_S^{microc} and $\rho_S^{\text{modapprox}}$ when the densities of states are as in (58).

In the spirit of the last paragraph of Section V we expect the leading term in S_S^{microc} to be given by

$$S_S^{\text{microc}} \simeq \int_0^E P_S(\epsilon) q\epsilon^2 d\epsilon \simeq \frac{qE^2}{2} \quad (69)$$

where, for the first approximate equality, we have replaced the logarithm in (44) by its ‘main part’ – i.e. by the exponent, $K\epsilon^2$, in the first equation in (58), and, for the second approximate equality, we have used the fact that the energy probability density, $P_S(\epsilon)$ (see (61) and Figure 4) consists of two sharp peaks, each of area 1/2, one located at $\epsilon = 0$ and one at $\epsilon = E$.

Proceeding similarly for $S_S^{\text{modapprox}}$, we similarly expect the leading term to be given by approximating (45) by

$$S_S^{\text{modapprox}} \simeq \int_0^{E/2} P_S(\epsilon) q\epsilon^2 d\epsilon + \int_{E/2}^E P_S(\epsilon) q(E-\epsilon)^2 d\epsilon \simeq 0. \quad (70)$$

It is straightforward to check that the error in both (69) and (70) is only of order 1 in E ; one needs only to be careful to realize that this is one situation in which (cf. Endnote [14]) it is important to work with the discrete sum versions, (40) and (41) or alternatively (42) and (43), of our entropy formulae; if one were to work unthinkingly with (44) and (45), one might (wrongly) conclude there is a (for some values of K , q and E , negative!) correction to both equations of form $\log(K/qE) + O(1)$ – the problem being caused by the steeply rising behaviour of $P_S(\epsilon)$ near $\epsilon = 0$ and $\epsilon = E$.

Thus, for our densities of states which grow as quadratic exponentials, there is an even more dramatic difference between the value of S_S^{microc} and the value of $S_S^{\text{modapprox}}$ than we found, in Section V, for densities of states which grow exponentially with energy (where they differed by a factor of 2).

VII. MORE ABOUT ENTROPY

Note: The reader may wish to skip this, and the next, section on a first reading and go directly to Section IX.

A. Special facts about the entropy when the energy probability density is sharply peaked

In the case of our power-law density of states example, an alternative way of arguing that the location of the peak of the energy probability density, $P_S(\epsilon)$, of the system, S, is given by the formula (31) is to assume foreknowledge of the existence of a (single) peak in the energy probability density, $P_S(\epsilon)$, in the interior of the energy-interval $[0, E]$ and then to note that, by (10), this must occur at an energy, ϵ for which

$$\frac{d \log \sigma_S(\epsilon)}{d\epsilon} + \frac{d \log \sigma_B(E - \epsilon)}{d\epsilon} = 0 \quad (71)$$

which easily implies (31).

In a popular approach (cf. also [43]) to such problems involving the microcanonical ensemble of a pair of weakly coupled systems with such a (say unique, interior) peak, such a calculation often appears in the following guise:

One writes $\epsilon = \epsilon_1$ and $E - \epsilon = \epsilon_2$. One calls $\log \sigma_S(\epsilon)$ “ $S_1(\epsilon_1)$ ”, and one calls $\log \sigma_B(E - \epsilon)$ “ $S_2(\epsilon_2)$ ”. Then one writes the equations

$$\epsilon_1 + \epsilon_2 = E,$$

$$\frac{\partial S_1}{\partial \epsilon_1} - \frac{\partial S_2}{\epsilon_2} = 0,$$

(equivalent to (71) and

$$\frac{\partial^2 S_1}{\partial \epsilon_1^2} + \frac{\partial^2 S_2}{\partial \epsilon_2^2} < 0 \quad (72)$$

(expressing the fact that it is a peak and not a trough).

It is often then assumed, or, at least, tacitly implied, that $S(\epsilon_1)$ and $S(\epsilon_2)$ are the “entropies” of Systems 1 and 2 (our systems S and ‘energy bath’ B) and that $\partial S_1/\partial \epsilon_1$ and $\partial S_2/\partial \epsilon_2$ are the “temperatures” of Systems 1 and 2. Finally, the equation (72) is interpreted as telling us that Systems 1 and 2 are in “stable equilibrium”.

Concerning this popular approach, we would remark and emphasize:

(a) $S(\epsilon_1)$ and $S(\epsilon_2)$ (our $\log \sigma_S(\epsilon)$ and $\log \sigma_B(E - \epsilon)$) are *not* entropies (they are logarithms of densities of states). To make sense of the logarithms one would, at least, need to multiply $\sigma_S(\epsilon)$ and $\sigma_B(E - \epsilon)$ by “constants” with the dimensions of energy, first, to make the overall arguments of the logarithms dimensionless. This may not matter if the resulting logarithm is anyway destined to be differentiated with respect to ϵ to define a ‘temperature’ (see Paragraph (b) below). However it *will* matter if one wishes to talk meaningfully about the logarithms themselves (evaluated at the peak values of ϵ and $E - \epsilon$) as ‘entropies’. One could, of course, insert, in each logarithm, an arbitrary constant with the dimensions of energy, and try to argue that it doesn’t make much difference, in practice, what is the precise value of this constant provided it is of a “reasonable” order of magnitude. But, even if it were the case that all that was at stake was such a “constant”, one would expect, in a fundamental understanding of the origin of entropy, its value to be determined in terms of the physical parameters of the problem. In fact, as we shall see below, what actually needs to be inserted is not a constant, but rather (for given system and energy-bath densities of states) a quantity (which we call Q below) with the dimensions of energy which (like the peak values of ϵ and $E - \epsilon$ themselves) depends on the totem energy, E .

(b) It is true that one can think of $\partial S_1/\partial \epsilon_1$ and $\partial S_2/\partial \epsilon_2$ as ‘energy-dependent temperatures’ in the sense (cf. Section I A and Endnote [1]) that, if System 1 had energy $\hat{\epsilon}_1$ and were uncoupled to System 2, but, rather, coupled to another and much smaller system, then that smaller system would likely get itself into a thermal state at the temperature $\partial S_1/\partial \epsilon_1$ evaluated at $\hat{\epsilon}_1$ (and similarly for System 2). However, in the ‘equilibrium’ in question, where System 1 and System 2 are coupled to one-another and neither can be regarded as ‘small’, *neither* System 1 nor System 2 is in a thermal state (as we have shown in Section II for our power-law case)!

(c) Finally, this ‘popular’ point of view is only of value in cases where the energy probability density (say of System 1) has a peak. Whereas, we emphasize, as explained in this paper, one still predicts definite energy probability density functions when System 1 and System 2 have densities of states (such as our equal exponential and equal quadratic exponential cases discussed in Sections III and VI) which do not lead to an interior peak. (In the equal exponential case, we find an energy probability density which is flat, and, in the quadratic exponential case, it is concave with peaks at the extremities of the range $[0, E]$ which are not ‘maxima’ in the sense of the above equations for S_1 and S_2 .) In such latter cases, the significant quantity of interest is not the location of a peak (there may even be no peak) but rather the full energy probability density function itself.

We turn next to the value of the entropy of the system in cases where, for given system and energy-bath densities of states, and given total energy, E , the energy probability density, $P_S(\epsilon)$, has a single peak, say at $\epsilon = \epsilon_0$. We are, of course, interested in both of the entropies, S_S^{microc} and $S_S^{\text{modapprox}}$. As we anticipated in Point (a) above, when the totem is in the microcanonical ensemble with energy in our interval $[E, E + \Delta]$, S_S^{microc} will be $\log(Q\sigma_S(\epsilon_0))$ for a suitable quantity, Q , with the dimensions of energy determined by the parameters of the problem, although we emphasize again that Q is not a “constant”; rather (for given system and energy-bath densities of states) it is a certain function of totem energy, E , which has the dimensions of energy and which is determined by the detailed shape of the peak in $P_S(\epsilon)$. In fact, by the general formula (44) (one can see that the issues mentioned in Endnote [14] will not be relevant for a sharp peak which is well inside the interior of the interval $[0, E]$) we will have

$$S_S^{\text{microc}} = \int_0^E P_S(\epsilon) \log(L\sigma_S(\epsilon)) d\epsilon - \int_0^E P_S(\epsilon) \log(LP_S(\epsilon)) d\epsilon$$

where we have *temporarily* introduced an arbitrary non-zero constant, L , with the dimensions of energy, which will of course cancel out in the final result.

Since $P_S(\epsilon)$ is, by assumption, sharply peaked at $\epsilon = \epsilon_0$, and assuming $\sigma(\epsilon)$ is relatively slowly varying (as will be true in typical examples such as the power-law density of states case treated below) the value of the first integral will be very well approximated by $\log(L\sigma_S(\epsilon_0))$. The second integral will obviously take the form $\log(L/Q)$ where Q is a quantity with the dimensions of energy which can in principle be computed in terms of our system and energy-bath densities of states and the value of E . So we will have

$$S_S^{\text{microc}} = \log(Q\sigma_S(\epsilon_0)). \quad (73)$$

On the other hand, if we consider the totem to be in a pure state, randomly chosen amongst all states with energy in the range $[E, E + \Delta]$, then, by (45), we expect the entropy to most probably be very close to $S_S^{\text{modapprox}}$ given by

$$S_S^{\text{modapprox}} = \int_0^{E_c} P_S(\epsilon) \log(L\sigma_S(\epsilon)) d\epsilon + \int_{E_c}^E P_S(\epsilon) \log(L\sigma_B(E - \epsilon)) d\epsilon - \int_0^E P_S(\epsilon) \log(LP_S(\epsilon)) d\epsilon$$

which, by a similar argument, and in view of the definition of E_c (see after Equation (15)), will be very well approximated by

$$S_S^{\text{modapprox}} = \min(\log(Q\sigma_S(\epsilon_0)), \log(Q\sigma_B(E - \epsilon_0)))$$

for the same value of Q .

We next illustrate the computation of Q in the case where both of system, S, and energy bath, B, have the power law densities of states (18) which we discussed in Section II. We have

$$\log(L/Q) = \int_0^E P_S(\epsilon) \log(LP_S(\epsilon)) d\epsilon$$

where $P_S(\epsilon)$ is given by (30) with γ as in (32). As long as N_S and N_B are comparable in size, the location, (31), of the peak, ϵ_0 , will be far from the extremities of the interval $[0, E]$ and we may clearly replace the limits of the above integral by $-\infty$ and ∞ with very little error.

So

$$\begin{aligned} \log(L/Q) &\simeq \frac{1}{E} \sqrt{\frac{\gamma}{\pi}} \int_{-\infty}^{\infty} \log\left(\frac{L}{E} \sqrt{\frac{\gamma}{\pi}} \exp\left(-\gamma \frac{(\epsilon - \epsilon_0)^2}{E^2}\right)\right) \exp\left(-\gamma \frac{(\epsilon - \epsilon_0)^2}{E^2}\right) d\epsilon \\ &= -\log\left(\frac{L}{E} \sqrt{\frac{\gamma}{\pi}}\right) + \frac{\gamma}{E^3} \sqrt{\frac{\gamma}{\pi}} \int_{-\infty}^{\infty} (\epsilon - \epsilon_0)^2 \exp\left(-\gamma \frac{(\epsilon - \epsilon_0)^2}{E^2}\right) d\epsilon. \end{aligned}$$

The second term above may easily be calculated by writing it as

$-(\gamma^{3/2}/E\pi^{1/2})\partial/\partial\gamma \int_{-\infty}^{\infty} \exp\left(-\gamma \frac{(\epsilon - \epsilon_0)^2}{E^2}\right) d\epsilon$. This is equal to $(\gamma^{3/2}/E\pi^{1/2})\partial/\partial\gamma(E\pi^{1/2}\gamma^{-1/2}) = 1/2$. So

$$\log(L/Q) \simeq -\log\left(\frac{L}{E} \sqrt{\frac{\gamma}{\epsilon\pi}}\right)$$

or, equivalently,

$$Q = \sqrt{\frac{e\pi}{\gamma}} E. \quad (74)$$

For the purposes of the comparison we make in Section II, we also calculate Q for a *thermal* state at some given inverse temperature, β , of a system, S , with a power law density of states $\sigma(\epsilon) = A\epsilon^N$. By (27) we will now have

$$\begin{aligned} \log(L/Q) &= \frac{\beta^{N+1}}{N!} \int_0^\infty \log\left(L \frac{\beta^{N+1}}{N!} \epsilon^N e^{-\beta\epsilon}\right) \epsilon^N e^{-\beta\epsilon} d\epsilon \\ &= -\log\left(\frac{L\beta}{N!}\right) - \frac{\beta^{N+1}}{N!} \int_0^\infty (N \log(\beta\epsilon) - \beta\epsilon) \epsilon^N e^{-\beta\epsilon} d\epsilon. \end{aligned}$$

We may do the integral here by noticing that $\int_0^\infty x^n \log x e^{-bx} dx = d/d\alpha \int_0^\infty x^\alpha e^{-bx} dx|_{\alpha=n} = d/d\alpha (b^{-(\alpha+1)} \Gamma(\alpha+1))|_{\alpha=n} = -(\log b) b^{-(\alpha+1)} \Gamma(\alpha+1) + b^{-(\alpha+1)} d/d\alpha \Gamma(\alpha+1)|_{\alpha=n} = -(\log b) b^{-(n+1)} n! + b^{-(n+1)} \Gamma(n+1) \psi(n+1)$ where Γ denotes the gamma function and $\psi(n+1)$ the digamma function (see e.g. [30]) of $n+1$ which (see again [30]) is equal to $\sum_{k=1}^n 1/k - C$ where C is Euler's constant ($= 0.5772\dots$). Using this, we conclude that $\log(L/Q) =$

$$-\log(L\beta) + \log N! - N\psi(N+1) + N + 1$$

which, using Stirling's approximation (which tells us that $\log n! = (n+1/2) \log n - n + (1/2) \log(2\pi) + 1/(12n) + O(1/n^2)$) and the asymptotic expansion of $\psi(n+1) = \log n + 1/(2n) + O(1/n^2)$ is equal to

$$-\log(L\beta) + \frac{1}{2} \log N + \frac{1}{2} \log(2\pi) + \frac{1}{2} + O(1/N) \quad (75)$$

from which we conclude that

$$Q \simeq \frac{\sqrt{2e\pi N}}{\beta}. \quad (76)$$

We note that if we identify N here with N_S and if $N_B \gg N_S$, then, by (32), $\gamma \simeq N_B^2/2N_S$. If, additionally, we take β in (76) to be $d \log \sigma_B(\epsilon)/d\epsilon|_{\epsilon=E}$ which (with $\sigma_B(\epsilon) = A\epsilon^{N_B}$) equals N_B/E , then the Q of (74) and (76) both take the same value $\sqrt{2e\pi N_S} E/N_B$. This agreement is to be expected since, as we discussed in Sections I and II, in this situation, where S is small and in the case of a total microcanonical state, the reduced density operator of S will be close to that of a thermal state at inverse temperature $d \log \sigma_B(\epsilon)/d\epsilon|_{\epsilon=E}$. So the agreement of the two Q in this regime serves as a check on the correctness of our two formulae (74) and (76).

However, in Section II we were interested in comparing the properties of the (as we show there) *non-thermal* reduced state of S when N_S and N_B are of comparable size and the totem is in a microcanonical state with the properties of a thermal state of S with the same expected energy. Treating, for simplicity, the case where $N_S = N_B = N$, say, the relevant inverse temperature, β , is $2(N+1)/E$ ($\simeq 2N/E$ for large N) and γ (32) is $4N$. We then find that the 'thermal' Q (Equation (76)) becomes $\sqrt{e\pi/2NE}$ whereas the 'microcanonical' Q (Equation (74)) becomes $(1/2)\sqrt{e\pi/NE}$. Thus the entropy of the comparison thermal state of S is bigger than the entropy, S_S^{microc} , of the reduced state of S by $\log 2/2$. While this is a small difference it is conceptually significant that it is not zero.

B. On the entropy of the totem and more about Δ

Our general framework involves a system, S , and an energy bath, B , comprising a totem. A natural question is: What is the relationship between the entropy of S , the entropy of B , and the entropy of the totem? The answer to this question depends, first of all, on whether we are contemplating the, traditional, microcanonical, scenario, or the modern scenario in which the state of the totem is pure – albeit chosen at random amongst the set of states in our energy range. In the latter, modern, scenario, there is only one entropy: As we mentioned in Sections I E and IV, the (von Neumann) entropy of S is equal to the (von Neumann) entropy of B and both are equal to the {system}-{energy bath} entanglement entropy of the totem; the von Neumann entropy of the totem is of course zero.

In the microcanonical scenario, one might, naively, expect the entropy of the totem to be the sum of the entropy of S and the entropy of B . But, as we shall see, this is not true. One way to see that it cannot be true is to notice (see again below) that the entropy of the totem (which will obviously have the value $\log M$, M as in (3) and (7)) – below

we shall call this $S_{\text{totem}}^{\text{microc}}$ – depends on the width, Δ , of our energy band $[E, E + \Delta]$ whereas, as we saw in Section IV (for a suitable range of Δ) the entropies of S and B do not. What is of course always true (and applies to both modern and microcanonical scenarios) is the property of *subadditivity* [31] which guarantees that the the entropy of the totem, must be less than or equal to the sum of the entropies of S and B.

Actually, it turns out, in all three cases we have studied here (i.e. with system and energy bath densities of states of power-law form and [equal] exponential or quadratic exponential form) that the sum of the entropies of S and B is *close to* the entropy of the totem. We have not attempted to formulate any general precise statement of what we mean by this, nor have we attempted to offer any general explanation as to why this should be the case but content ourselves simply with making the content of this statement manifest for each of our three density-of-states models:

For our (equal) exponential densities of states (34) we notice that, by (35)

$$S_{\text{totem}}^{\text{microc}} = \log M = bE + \log(c^2 E \Delta)$$

whereas, by (54)

$$S_S^{\text{microc}} + S_B^{\text{microc}} = bE + \log(c^2 E^2).$$

For our (equal) quadratic exponential densities of states (58), by (62)

$$S_{\text{totem}}^{\text{microc}} = \log M = qE^2 + \log\left(\frac{K^2 \Delta}{2E}\right)$$

whereas (see (69) and the paragraph after (70))

$$S_S^{\text{microc}} + S_B^{\text{microc}} = qE^2 + O(1).$$

For our power-law densities of states (18), we first obtain a good approximate formula for M by noting that the value of the integral in (19) is equal to the ratio of the maximum value of its integrand, $\epsilon_0^{N_S}(E - \epsilon_0)^{N_B}$ (ϵ_0 as in (31)) to its maximum value when normalized, which, by our Gaussian approximation, (30), is well-approximated by $(1/E)\sqrt{\gamma/E}$, γ as in (32). Thus we have (to a very good approximation)

$$S_{\text{totem}}^{\text{microc}} = \log M = \log\left(A_S A_B E \Delta \sqrt{\frac{\pi}{\gamma}} \epsilon_0^{N_S} (E - \epsilon_0)^{N_B}\right)$$

whereas, by (73) for S and its obvious counterpart for B and (74),

$$S_S^{\text{microc}} + S_B^{\text{microc}} = \log\left(A_S A_B E^2 \frac{e\pi}{\gamma} \epsilon_0^{N_S} (E - \epsilon_0)^{N_B}\right).$$

We see that subadditivity in the exponential case entails $\Delta < E$. In the quadratic exponential case, (and neglecting the $O(1)$ term) it entails $\Delta < 2E/K^2$, and in the power-law case, it entails $\Delta < e\sqrt{\pi/\gamma}E$. When $N_S = N_B = N$, say, $\gamma = 4N$ and this latter inequality amounts to $\Delta < (e\sqrt{\pi}/2)(E/\sqrt{N})$. The first of these inequalities ($\Delta < E$) is obviously consistent with almost any sort of smallness assumption on Δ . The other two inequalities indicate a need to be more precise than we were, in our rather sketchy remarks in Section I and in our subsequent derivations, about what is the appropriate range of Δ for any given pair of densities of states, σ_S and σ_B , in order for our arguments and approximations to be valid. We shall, however, not pursue this issue further in the present paper except to deduce that the above inequalities must be necessary conditions on the value of Δ .

VIII. MORE ABOUT THERMALITY: PURIFICATION

Throughout the preceding sections we have assumed (see the paragraph after Equation (2)) that both our system, S, and our energy bath, B, have densities of states which are positively supported (i.e. the Hamiltonians H_S and H_B are positive) and monotonically increasing and we have been concerned exclusively with totem states which are (close to) stationary states for a totem Hamiltonian, (2), which weakly couples S and B. In this short section, we point out that the prospects for the thermality of either S or B become much less constrained if we relax some of these assumptions. In particular, and in the spirit of the ‘modern’ approach, given *any* system density of states, $\sigma_S(\epsilon)$, whatsoever (provided only it grows sufficiently slowly for the desired thermal state to exist) one can always find an energy bath density of states and a pure totem state such that the reduced state of the system is an exactly thermal state at any given temperature.

In fact, there is a well-known procedure, in the spirit of the modern approach, known as ‘purification’ (see e.g. [32] and references therein; see also the papers on ‘thermofield dynamics’ [34] and [35] which are based on the same idea – we note that the term ‘purification’ seems to be due to Powers and Størmer [33]) by which a system, S , with any density of states whatsoever (but we shall assume it to be positively supported and monotonically increasing) and in any non-pure state one wishes to prescribe (but we are interested in a thermal state at some inverse temperature β) may be provided with a notional energy bath, B , such that there is a choice of pure state on the resulting totem for which the reduced density operator on S is equal to the prescribed state.

The essential idea of purification is based on the fact that any density operator, ρ , on a Hilbert space, \mathcal{H} , takes the form

$$\rho = \sum_i \rho_i |\psi_i\rangle\langle\psi_i|,$$

the ρ_i being positive numbers which sum to 1 and ψ_1, ψ_2, \dots being an orthonormal basis for \mathcal{H} , and on the observation that this can be viewed as arising as the partial trace over the second copy of \mathcal{H} , in the tensor product $\mathcal{H} \otimes \mathcal{H}$, of the pure density operator, $|\Psi\rangle\langle\Psi|$, where

$$\Psi = \sum_i \rho_i^{1/2} \psi_i \otimes \psi_i.$$

The easiest way to see this is to notice that, for any linear operator, A , on \mathcal{H} ,

$$\langle\Psi|(A \otimes I)\Psi\rangle_{\mathcal{H} \otimes \mathcal{H}} = \text{tr}(\rho A)_{\mathcal{H}}.$$

If we now specialize to the thermal situation where the ψ_i are the energy eigenstates of a Hamiltonian, H , on \mathcal{H} with energy eigenvalues say e_i and $\rho_i = Z^{-1}e^{-\beta e_i}$ and identify \mathcal{H} both with the system Hilbert space, \mathcal{H}_S , and also with the Hilbert space, \mathcal{H}_B , of our notional energy bath, then, if we also take the energy bath Hamiltonian to equal H (and therefore the energy levels of the energy bath to be the same as the energy levels of the system) then Ψ provides us with a pure totem state with the property that the reduced state of the system (and also the reduced state of the energy bath) is exactly thermal. With the notation of Section I A, we would write $\mathcal{H}_B = \mathcal{H}_S$ and take Ψ on $\mathcal{H}_S \otimes \mathcal{H}_B$ to be given by

$$\Psi = Z_{S,\beta}^{-1/2} \sum_{\epsilon=\Delta}^{\infty} e^{-\beta\epsilon/2} \sum_{i=1}^{n_S(\epsilon)} |\epsilon, i\rangle \otimes |\epsilon, i\rangle. \quad (77)$$

Then the partial trace of $|\Psi\rangle\langle\Psi|$ over \mathcal{H}_S will equal the $\rho_{S,\beta}^{\text{Gibbs}}$ of Equation (13).

This achievement of exact thermality contrasts with the situation discussed in Section I D (see in and after the paragraph containing Equation (17)) where (on the ‘modern approach’) the totem state is assumed to be randomly chosen from amongst totem states in a narrow range of energies for a weakly coupled total Hamiltonian. As we saw there and in the rest of Part 1, with that assumption, and for systems of comparable size, thermality can never be achieved exactly and can only be approximately achieved for certain special densities of states – such as, in particular, the exponential and quadratic exponential cases discussed in Sections III and VI. However, in the purification mechanism described here, the state, Ψ , of the totem is – say if we regard the totem Hamiltonian to be given by Equation (2) with $H_B = H_S$ (and, say, with no coupling term at all) – clearly not even close to an energy eigenstate; in other words, the totem is in a highly non-stationary state. We remark that this purification mechanism does not seem to play much of a role in everyday physics as a mechanism by which a system can get to be hot, although, interestingly, essentially this mechanism has been made use of in the laboratory [36] to produce thermal states of photons. (See also the remark about the Unruh effect in the next paragraph and the remarks about quantum black holes in Section IX.)

We further remark that there is an alternative reinterpretation of Equation (77) in which one ascribes to the ‘energy bath’, B , the Hamiltonian $H_B = -H_S$ (and substitutes these Hamiltonians into the totem Hamiltonian formula (2)). With this interpretation, the state of the totem is an eigenstate of totem energy (with totem energy eigenvalue zero!) and so a stationary state. But now the density of states of the energy bath is negatively supported! This latter interpretation can be said to be realized in the Unruh effect (see e.g. [37] and reference therein) whereby the vacuum state of a relativistic quantum field theory in Minkowski space, restricted to a right-Rindler wedge, is a thermal state with respect to Lorentz boosts; the left-Rindler wedge plays the role of our energy bath, B , and this can be thought of as a copy of the right-wedge quantum system but with a Hamiltonian which is the negative of the right-wedge Rindler Hamiltonian.

IX. IMPLICATIONS FOR QUANTUM BLACK HOLES

A. The (problematic) connection between our results on quadratic exponential densities of states in Section VI and black hole thermodynamics

There is a striking, at least superficial, resemblance between Equations (66) and (69) in Section VI – i.e. $\beta = 2qE$ and $S_S^{\text{microc}} (= S_B^{\text{microc}}) = qE^2/2$ (up to an $O(1)$ correction) – for the inverse temperature and the traditional microcanonical entropy of a weakly coupled system, S, and energy bath, B, with equal quadratic exponential densities of states (58) and the equations (see Section IB) $\beta = 8\pi G\mathcal{M}$ and $S = 4\pi G\mathcal{M}^2$ for the Hawking inverse temperature and entropy of a Schwarzschild black hole. In fact, if we identify, say (see below), the mean energy of B – i.e. (see (63)) half the totem mean energy, $E/2$ – with the black hole mass, \mathcal{M} , and identify q with $2\pi G$, the entropy, S_B^{microc} matches the Hawking entropy and the (inverse) temperatures match too.

This might seem to suggest that, if one identifies the system, S, with ‘matter’ and the energy bath, B, with ‘gravity’, then the traditional microcanonical strand of Section VI may provide a good model for a black hole in contact with its thermal atmosphere in a box and a good explanation for the microscopic origin of the entropy of this system. (Of course, we must bear in mind that, in this model, the state is only thermal in the approximate sense explained in Section VI.) And on the other hand, our result, (70), that, with the densities of states (58), the ‘modern’ totem pure state entropy, $S_S^{\text{modapprox}}$ vanishes (up to a term of order 1 in E) might seem to be at odds with our matter-gravity entanglement hypothesis described in [17–19] and in Section IB – which entails that the total matter-gravity state of a black hole is a pure state. However, we need to realize that the results of Section VI assume that the dynamics of the totem is governed by a totem Hamiltonian of the schematic form (2) with both S and B (now to be interpreted as ‘matter’ and ‘gravity’) Hamiltonians positive and weakly coupled. Yet, notoriously, it seems unlikely to be possible to have a quantum theory of gravity within the scope of these basic assumptions (albeit these assumptions do seem to apply to the weak-string coupling limit if the fundamental degrees of freedom are taken to be those of a string rather than of the gravitational field itself – see Section IX B). Already classical general relativity is nonlinear and, unlike the situation for (2), energy (mass) is not additive. So the fact that the mean energies of system and energy bath (modelling matter and gravity) are equal in our model seems strange. (It is also unclear whether we should identify the entropy of the black hole with S_B^{microc} as we did above, or with $S_B^{\text{microc}} + S_S^{\text{microc}}$, which is twice as big, or with $\log M$ – see also Section VII B.) Furthermore (see the remarks after Equation (63)) in this model, the mean energy, $E/2$, of each of matter and of gravity is anyway just the mean of an energy probability density (the $P_S(\epsilon)$ of (68) and of Figure 4) which is peaked at the extremes, $\epsilon = 0$ and $\epsilon = E$, while of course (cf. after Equation (63)) the energy probability densities of matter and gravity are perfectly anticorrelated. So the model predicts large statistical fluctuations, with (to the extent that it makes sense to talk about the energy of subsystems in general relativity) probability distributions for matter and gravity being such that, approximately, with probability one half, gravity has all the energy and matter none, and with probability one half, matter has all the energy and gravity none. The latter case (where there is presumably no black hole) is then a particular problem because (see the next paragraph) presumably the quadratic exponential form of the density of states presupposed in the model for matter becomes invalid when a black hole is not present.

In fact, turning to more specifically quantum aspects, aside from all the usual problems of quantum gravity (non-renormalizability etc.) it would seem to be incorrect to assume that one can ascribe a single density of states to each of gravity and matter throughout changes of state which include the formation of black holes. Rather it would seem that one has to assign, in some sense, a ‘state-dependent density of states’ to matter; in the absence of black holes, the densities of states of common forms of matter (including photons) grow much more slowly than quadratic exponentials, while, plausibly, when a black hole is present, they do grow as quadratic exponentials (with subleading corrections). (Evidence for this latter statement is provided by the ‘brick wall’ approach [50–52] which suggests that the matter entropy is comparable to the gravitational entropy when a black hole is present. We shall also argue in [23] that the string scenario we advocate in Section IX B and discuss further in [22, 23] leads, plausibly, to just such a state-dependent density of states for matter.) Moreover, the situation is further complicated by the very different status of the concept of ‘time’ in general relativity from that presupposed in traditional formulations of quantum theory.

In the light of all these problems and difficulties, and of our current lack of knowledge as to how to resolve them (other than to assume that a black hole is an [ill-understood] strong string-coupling limit of a certain [better understood] state of string theory at weak coupling – see Section IX B) it seems to us still reasonable to cling to our matter-gravity entanglement hypothesis and our entanglement picture of black hole equilibrium (see Section IB). Indeed, there would seem just as much reason to believe in a model along the following ‘modern’ lines (inspired by the idea of ‘purification’ outlined in Section VIII) as to believe in the above model within the traditional microcanonical strand of Section VI:

A tentative possible ‘modern’ model with the correct Hawking value for the entropy: ‘Matter’ is modelled as a

‘system’, ‘S’, and gravity as an ‘energy bath’, ‘B’, which each have a density of states as in (58), and the total state of the matter-gravity totem which corresponds to a black hole of mass $\mathcal{M} = E/2$ is modelled by the pure density operator $|\Psi\rangle\langle\Psi|$ where Ψ is

$$\frac{1}{\sqrt{M}} \sum_{\epsilon=\Delta}^E \sqrt{n_B(E-\epsilon)} \sum_{i=0}^{n_S(\epsilon)} |\epsilon, i\rangle_S |\epsilon, i\rangle_B \quad (78)$$

where (cf. (4), (5) and (58)) $n_B(\epsilon) = \exp(q\epsilon^2)\Delta$ and $q = 2\pi G$.

This state is easily seen to have partial traces over S and B identical to those, i.e. the ρ_B^{microc} and ρ_S^{microc} of (11), of the microcanonical state (6) ρ_{microc} for the same densities of states (i.e. (58)) and thus, obviously, it will equally well predict an inverse temperature of $2qE$ and a system entropy (= energy bath entropy) of $qE^2/2$ (plus the same $O(1)$ correction).

While we have argued that this latter ‘modern’ model is no less justified than our above microcanonical model, in view of the difficulties and problems mentioned above, it is still quite unclear what status should be assigned to it or how seriously it should be taken. There is also an apparent flaw in this tentative model in that the pure state of the totem is far from being an energy-eigenstate. It could possibly be that this is the best one can do when one attempts to force a strong-coupling situation into a weak-coupling mould, or maybe the model should be modified along the lines of the alternative reinterpretation of Equation (77) in Section VIII so that the totem state is modelled as an energy eigenstate, at the expense of having densities of states which are not monotonically increasing and/or not positively supported. Finally, there is the same strange feature that we raised above for our microcanonical model, that both the gravity and the matter are modelled as having, on average, exactly half of the mass (i.e. $E/2$) of the totem. Also, as in the microcanonical model, the energy probability densities of matter and gravity are predicted to each have the same energy probability density (the $P_S(\epsilon)$ of (68) and of Figure 4) with equal-sized peaks at $\epsilon = 0$ and $\epsilon = E$. Albeit, interestingly (and related to the fact that the totem state is far from being an energy eigenstate) this model differs from the microcanonical model in that the two energy probability densities are now no longer anticorrelated but, instead, perfectly correlated: One sees immediately from (78) that when gravity has energy near 0, so will matter, and when gravity has energy near E , so will matter. So at least one of the problems we mentioned above (the one we referred to as a “particular problem”) for the microcanonical model seems to be alleviated in the above proposed ‘modern’ model. Another problem which is alleviated with this tentative modern model is that it is clear, in this modern model that the entropy should be identified with S_B^{modern} , whereas, as we remarked above, in the microcanonical model it was not clear whether it should be identified with S_B^{microc} or with (approximately – see Section VII B) twice this value; in the modern model, there *is* only one entropy – i.e. the S-B entanglement entropy ($= S_B^{\text{modern}} = S_S^{\text{modern}} \simeq S_B^{\text{modapprox}} = S_S^{\text{modapprox}}$)!

One would also wish to be able to relate our discussion, in Section VII A, of cases where the probability density is sharply peaked, to the results, [43], of Hawking on his microcanonical approach to quantum black holes. These latter results of Hawking do seem to form a physically compelling and coherent picture and one would like to understand whether and, if so, how, they can be reconciled with the modern strand of results in Section VII A even though they seem, superficially, to be more easy to understand in terms of the microcanonical strand of work there and seem, superficially, to be at odds with the modern strand. We hope to address this question elsewhere. Suffice it to say here that, again, the difficulties and problems mentioned above are of at least equal relevance also to this issue and thus it seems difficult, also for this issue, to reach a fully convincing conclusion either way [44].

B. Towards a better understanding of black hole entropy in terms of string entropy

Where we have been able to make a, we think, persuasive, case for the relevance, of the ‘modern’ strand of ideas in the present paper – used in combination with the (see Section I B) matter-gravity entanglement hypothesis of [17–19] – to the understanding of black hole entropy is with a model which relates the work in Sections III and V here, concerning densities of states which grow exponentially with energy, to an understanding of black hole entropy based on the idea that black holes are strong string-coupling limits of states of weakly coupled string theory. This application of our work to quantum black holes seems to be more well-founded because, unlike in the situations discussed above, we *would* expect the general assumptions we made at the outset here (positive Hamiltonians, weak coupling etc.) to be applicable to the weak-coupling regime of string theory.

In 1993, Susskind [38] proposed, and in 1997, Horowitz and Polchinski [40, 41], gave further evidence, of a semi-qualitative nature, for, the hypothesis that a (say 4-dimensional, Schwarzschild) black hole can be interpreted as the strong string-coupling limit of a certain state of string theory at weak coupling consisting of a (single) long string. These authors argued that one obtains, with this interpretation, an explanation of black hole entropy in terms of

known formulae, based on the ‘counting of states’, in string theory at low string coupling. Moreover, in related work on extremal, and near extremal, black holes (an important early paper was Strominger and Vafa [39]), full quantitative agreement was found between the results of such a string theory approach to black hole entropy (and other related quantities) and the previously established Hawking formulae. It was then claimed that this work, not only gave a microscopic explanation for black hole entropy but also, in view of the fact that string theory is a standard quantum mechanical theory with a unitary time evolution, that it resolved the Information Loss Puzzle (see Section IB). We next wish to argue that, while we agree all this work seems to provide us with an important clue towards the microscopic understanding of black-hole entropy which, plausibly, may well turn out to be consistent with a resolution to the Information Loss Puzzle, it cannot, by itself, be regarded as a complete explanation of these things. (We shall expand on these arguments in two companion papers, [22] and [23].) Our point is simply that, what is actually calculated in the cited work (for example in [39]) is not the *entropy* of a particular black hole state, but rather the (logarithm of the) *degeneracy* of a given black hole energy level. No explanation was given in the cited work as to why the logarithm of this degeneracy should be interpreted physically as an entropy. After all the n ’th energy level of the textbook non-relativistic Hydrogen-atom Hamiltonian has a degeneracy of n^2 but we would not predict from this that a Hydrogen atom has an entropy of $\log n^2$! Of course it is true that the logarithm of the degeneracy of an energy level is the same thing as the von Neumann entropy of the microcanonical density operator $(1/d) \sum_{i=1}^d |i\rangle\langle i|$ where we denote by $|i\rangle$, ($i = 1 \dots d$) the elements of a basis of states with the given energy. But if we were to attempt to interpret e.g. the Strominger Vafa results as meaning that a black hole should be modelled by such an (impure!) microcanonical state, then the Information Loss Puzzle would surely return: How, in a string theoretic description of a dynamical process of black hole formation, can a presumably pure initial string theory state evolve into such an (impure) microcanonical state? (Such a microcanonical state also wouldn’t fit with our picture of black holes as thermal states.)

Actually, the Horowitz-Polchinski work is couched in terms, not of the degeneracy of a particular energy level of string theory, but rather of the (averaged out) density of states of a long string. The problem is then compounded by the fact that a density of states is not a dimensionless quantity, so it is not physically meaningful to take its logarithm [45]. (In fact similar remarks apply to those we make in Paragraph (a) of Section VII A.)

Focusing on this Horowitz-Polchinski work, we shall next propose a modified version of their scenario, based on the modern strand of Sections III and V of the present paper, which seems to overcome the above difficulties and to offer the promise of a fully satisfactory explanation of black hole entropy in terms of string theory, consistent with unitarity and consistent with a resolution to the Information Loss Puzzle – namely with the resolution to the Information Loss Puzzle we proposed in [17–19] based on our matter-gravity entanglement hypothesis (see above and Section IB). This will be further discussed in [22] and further developed in [23], whose content we indicate very briefly at the end of this subsection.

We begin by briefly recalling the basic argument of Susskind, Horowitz and Polchinski [38, 40] as expounded in [41]. Their basic hypothesis is that, as one scales the string length scale, ℓ , up and the string coupling constant, g , down from their physical values, keeping Newton’s gravitational constant, $G = g^2 \ell^2$, fixed, a (4-dimensional) Schwarzschild black hole of mass \mathcal{M} will turn into a long string with roughly the same energy, $\epsilon = \mathcal{M}$. The density of states of such a long string, in the limit of weak coupling, is known, very roughly (i.e. omitting an approximately inverse-power prefactor – see below) to take the exponential form

$$\sigma_{\text{long string}}(\epsilon) = C_{\text{ls}} e^{\ell \epsilon} \quad (79)$$

(C_{ls} a constant with the dimensions of inverse energy of the same order of magnitude as ℓ).

The gist of the argument is that the ‘logarithm’ of this is approximately given by

$$S_{\text{long string}} = \ell \epsilon \quad (80)$$

and they refer to this quantity as (approximately) the ‘entropy of the long string at energy ϵ ’. They then argue that this should be equated with the entropy of a (Schwarzschild) black hole provided that one does the equating (i.e. during the process of scaling ℓ described above) when, to within an order of magnitude or so,

$$\ell = GM \quad (81)$$

which is roughly the ‘size’ of the black-hole. (Cf. the fact that the Schwarzschild radius is $2GM$.)

Combining (80) and (81) (and replacing ϵ by \mathcal{M}) they thus claim to predict that the entropy of the black hole will be within an order of magnitude or so of a constant times GM^2 which agrees, up to an undetermined value for the constant, with the Hawking value, $4\pi GM^2$, for the entropy of a black hole.

In our view, what one is actually entitled to say, instead, is that, the number of energy eigenstates of a black hole in a band of width Δ around energy ϵ will, by (79) be $\ell \epsilon + \log(C_{\text{ls}} \Delta)$, which, for a ‘reasonable-sized’ Δ will

be approximately the same as $\ell\epsilon$. The argument in the previous two paragraphs then tells us that the number of energy eigenstates of a black hole in a band of width Δ around energy (i.e. mass) \mathcal{M} will be within an undetermined constant, say C , of the order of 1, times $G\mathcal{M}^2$ (and thus, by the way, that the density of states of a Schwarzschild black hole should behave roughly as a constant times $\exp(CG\epsilon^2)$.) However, in our view, it remains a challenge to explain why the logarithm of this formula for the density of states of a black hole should equal (to within an unknown constant, C of the order of 1) the Hawking formula for black hole entropy.

In our attempt to meet this challenge, we first posit that the scenario in which a black hole goes over to a single long string should be replaced by a scenario in which an equilibrium state (i.e. energy eigenstate) consisting of a black hole in contact with its matter atmosphere in a suitable box (on our view described by a pure total state – see our ‘entanglement picture of black hole equilibrium’ in Section IB) with approximate total energy, E , goes over (again, as one scales the string length scale, ℓ , up and the string coupling constant, g , down from their physical values, keeping Newton’s gravitational constant, $G = g^2\ell^2$ fixed) to a (pure) equilibrium state, with a similar total energy, consisting of a single long string in contact with an atmosphere of small strings in a suitably rescaled box.

We now assume that the density of states of the long string takes (to the same rough approximation as above) the form of (79) and that the density of states of the stringy atmosphere, $\sigma_{\text{string atmosphere}}(\epsilon)$, takes (again, to a rough approximation) the similar, exponential, form

$$\sigma_{\text{string atmosphere}}(\epsilon) = C_{\text{sa}} e^{\ell\epsilon}. \quad (82)$$

If we now regard (most of) the stringy atmosphere as corresponding to ‘matter’ and as playing the role of our ‘system’, S, and the long string as corresponding to (most of) ‘gravity’ and as playing the role of our energy-bath, ‘B’, then it is plausible that these may be described by Hilbert spaces and Hamiltonians which, since we are at weak string coupling, should fall within the scope of the present paper, and in particular the dynamics should be described by a totem Hamiltonian of form (2). In view of the exponential growth of the densities of states, (79) and (82), we may therefore apply the formalism of Sections III and V (modified as explained in Endnote [29] to take into account the different prefactors, C_{sa} and C_{ls}). In particular, the modern strand of these sections tells us that a typical pure equilibrium state of our {string atmosphere}-{long string} totem with energy around E will, with a high probability, have an entropy very close to that given by Equation (55) with $b = \ell$ (with the modification to the logarithmic term given in Endnote [29]). I.e. ignoring the logarithmic term, by

$$S = \ell E/4,$$

while the (expected) energy of the long string, $\bar{\epsilon}_{\text{long string}}$ will (see again Endnote [29]) be given by

$$\bar{\epsilon}_{\text{long string}} = E/2$$

(and, of course the mean energy of the stringy atmosphere will also be $E/2$ in this model).

In parallel to the philosophy of [38, 40, 41], we now assume that, when we scale g back up and ℓ down, keeping $G = g^2\ell^2$ constant and keeping Ψ the ‘same’, we can equate $\bar{\epsilon}_{\text{long string}}$ with the mass, \mathcal{M} , of the black hole when $\ell = XG\mathcal{M}$, say, where X is an adjustable parameter of the order of 1. We thereby obtain the prediction $S = XGM^2/2$, as the value for the entanglement entropy of black hole and thermal atmosphere in the ‘same’ (i.e. after rescaling) state Ψ . But it is also plausible (as indicated above – see the relevant Endnote in [23] for further discussion) that this is approximately the same as the entanglement entropy between gravity and matter which, according to the matter-gravity entanglement hypothesis of [17–19] and Section IB is the physical entropy of the black hole. We thus predict that the physical entropy of our black hole is (approximately) $XGM^2/2$. This agrees with the Hawking entropy of $4\pi\mathcal{M}^2$ if we take $X = 8\pi$.

Furthermore, we showed, in Section III, that both S and B will be ‘ E -approximately semi-thermal’, in the sense explained there, at inverse temperature $\beta = \ell$. Equating this with the inverse black hole temperature when $\ell = XG\mathcal{M}$ predicts a black hole inverse temperature of $XG\mathcal{M}$ which, intriguingly, agrees with the inverse Hawking temperature for the same value of X (i.e. 8π). We remark that we would not have correctly predicted both Hawking temperature and Hawking entropy for a single value of X had we followed the traditional microcanonical, instead of the modern strand, of Section III nor if we had adopted the approach of [38, 40, 41] and defined the inverse temperature by $\beta = d(\log(\sigma_{\text{long string}}(\epsilon))/d\epsilon$. (In each case, the necessary values of X for fitting the Hawking entropy and the Hawking temperature would have differed by a factor of 2.) However we caution that it is not clear whether this nice feature of our modern model with exponential densities of states survives when (see next paragraph) we improve the model to include the appropriate approximately inverse-power prefactors. We discuss this further in [23]. Nevertheless, our main point is that a ‘modern’ model for black hole entropy, based on our matter-gravity entanglement hypothesis seems able to predict a temperature of the order of the Hawking temperature and an entropy of the order of the Hawking entropy.

The main deficiency in the above scenario is the adoption of the equations (79) and (82) for the approximate forms of the long-string and stringy-atmosphere densities of states. These formulae omit important (dimension-dependent) approximately inverse-power prefactors, and when one takes these into proper account, it turns out (with some, seemingly reasonable, assumptions) that the account of the origin of black hole entropy above and in [22] needs significant changes and is even, in certain respects, misleading, although one arrives at similar final conclusions. The prefactors are also needed to explain why an equilibrium weakly coupled string state in a box consists of a single long string surrounded by an atmosphere of short strings as we posited above. Also, the statistical spread in energy of the string (and hence the predicted statistical spread in energy of the black hole) around the mean energy $E/2$ will be altered with the correct prefactors. All these matters will be discussed in our second companion paper [23].

Part 2: Full explanation of the formula (15) and arguments for the validity of the proposition in Section ID

X. THE WORK OF LUBKIN AND PAGE AND OTHER PRELIMINARIES AND OUTLINE OF THE REMAINDER OF PART 2

In Section I, we mentioned the work of Lubkin, [10], where it is shown that a randomly chosen pure density operator, $\rho^{mn} = |\Psi\rangle\langle\Psi|$, on the tensor-product Hilbert space, $\mathcal{H}_m \otimes \mathcal{H}_n$, of a pair of quantum systems – \mathcal{H}_m being m -dimensional and \mathcal{H}_n being n -dimensional – will, for fixed m and $n \gg m$, have, with high probability, a reduced density operator, ρ_m^{mn} , on \mathcal{H}_m , which is close to the maximally mixed density operator – with components, in any Hilbert space basis, $\text{diag}(1/m, \dots, 1/m)$. We first need to recall some more details about this work as well as some further related developments which will be relevant throughout Part 2.

Lubkin justified the above statement and made it precise by obtaining a result which is (easily seen to be) equivalent to the following exact formula for the mean value (i.e. over Haar measure on the set of unit vectors Ψ) $\langle \text{tr}((\rho_m^{mn})^2) \rangle$, of $(\rho_m^{mn})^2$: In our notation

$$\langle \text{tr}((\rho_m^{mn})^2) \rangle = \frac{m+n}{mn+1}. \quad (83)$$

We shall re-derive this result of Lubkin with a somewhat different method in the next section (Section XI) since our full explanation of the formula in Equation (15) and our arguments for the validity of our proposition of Section ID will be closely based on it. Lubkin then gave a simple general argument which amounts to the statement that, for any density operator, ρ_m , on an m -dimensional Hilbert space, whenever $m\langle \text{tr}(\rho_m^2) \rangle - 1 \ll 1$, then the mean value, $\langle S(\rho_m) \rangle$, of the von Neumann entropy, $S(\rho_m)$, of ρ_m will be well-approximated by

$$\langle S(\rho_m) \rangle \simeq \log m - \frac{1}{2} (m\langle \text{tr}(\rho_m^2) \rangle - 1). \quad (84)$$

Applying this result to ρ_m^{mn} , (83) implies that whenever $m \ll n$,

$$\langle S(\rho_m^{mn}) \rangle \simeq \log m - \frac{m^2 - 1}{2(mn + 1)} \quad (85)$$

which may also be regarded as an alternative quantitative expression of the qualitative property that, when $m \ll n$, most ρ_m^{mn} must be close to maximally mixed.

In (84) and (85) above, the von Neumann entropy is defined in the usual way as in Equation (16).

Around 15 years later, Page [46] showed that the formula

$$\langle S(\rho_m^{mn}) \rangle \simeq \log m - \frac{m}{2n} \quad (86)$$

is a good approximation (with error term of order $1/mn$) whenever $1 \ll m \leq n$, and noted that this agrees with (85) on their common domain of validity [47]. We note here, in passing that, combining the two estimates (85) and (86), we can clearly write, simply, that whenever $m \leq n$, $\langle S(\rho_m^{mn}) \rangle = \log m - m/2n + O(1/mn)$ (since, when m and n are both of order 1, the entropy can anyway only be of order 1) and hence, combining this result with m and n

interchanged with the equality of $S(\rho_m^{mn})$ and $S(\rho_n^{mn})$ [48] (by which we mean the reduced density operator of ρ^{mn} on \mathcal{H}_n – see Section XI)

$$\langle S(\rho_m^{mn}) \rangle = \log(\min(m, n)) - \frac{\min(m, n)}{2\max(m, n)} + O\left(\frac{1}{mn}\right). \quad (87)$$

In the remainder of Part 2, we first introduce, in Section XI, a useful coordinatization for unit vectors in an N -dimensional Hilbert space in terms of which the ‘Haar’ measure of Section I takes a particularly convenient form. In order to prepare the ground for our subsequent generalization (see below) we then use this coordinatization to obtain an alternative derivation of Lubkin’s result (83). We also discuss in more detail the qualitative consequence of Lubkin’s result concerning the ‘almost maximal mixedness’ of the density operator ρ_m^{mn} when $m \ll n$ (as previously pointed out by Lubkin, as mentioned in Section I) and we also point out a related important second qualitative consequence concerning the nature of ρ_m^{mn} in the ‘opposite’ situation when $m \gg n$. Then, in Section XII, we use a generalization of our alternative derivation of Equation (83) as well as a suitable generalization of our argument for its two qualitative consequences to give the full statement of Equation (15) including an explanation of how the $n_B(E - \epsilon)$ -dimensional subspace of the $(n_S(\epsilon)$ -dimensional) energy- ϵ subspace of \mathcal{H}_S spanned by the $|\widetilde{\epsilon}, i\rangle$ depend on Ψ and also to give our argument for the validity of our proposition (stated in full in Section ID) that ρ_S^{modern} (see the discussion after (15)) is well approximated by the $\rho_S^{\text{modapprox}}$ of Equation (15). We end, in Section XIII, with two calculations which provide confirmatory evidence of the goodness of our approximation in situations such as those we discuss in Part 1.

XI. A USEFUL REPRESENTATION OF HAAR MEASURE AND DETAILS ON, AND FURTHER CONSEQUENCES OF, LUBKIN’S RESULT

Let \mathcal{H} be an N -dimensional Hilbert space and let $\{E_1 \dots E_N\}$ be an arbitrary orthonormal basis. Then, as usual, we coordinatize an arbitrary vector, $\psi \in \mathcal{H}$, by the N -tuple of complex numbers (z_1, \dots, z_N) where $\psi = \sum_{a=1}^N z_a E_a$. ψ is, of course, then a unit vector if and only if $\sum_{a=1}^N |z_a|^2 = 1$. So the set of normalized vectors in our Hilbert space is coordinatized as the unit sphere in \mathbb{C}^N . (Writing $z_a = x_a + iy_a$ etc. we see that this is obviously the ‘same thing’ as the real unit $(2N-1)$ -sphere.) Next we change to polar coordinates in each copy of \mathbb{C} by setting $z_a = r_a e^{i\theta_a}$, whereupon the usual volume element $dz_1 \dots dz_N$ on \mathbb{C}^N takes the form $r_1 \dots r_N dr_1 \dots dr_N d\theta_1 \dots d\theta_N$. Changing coordinates further from $(r_1, \dots, r_N; \theta_1, \dots, \theta_N)$ to $(r_1, \dots, r_{N-1}; R; \theta_1, \dots, \theta_N)$, where

$$R^2 = \sum_{a=1}^N r_a^2, \quad (88)$$

this volume element is easily seen to become $r_1 \dots r_{N-1} R dr_1 \dots dr_{N-1} dR d\theta_1 \dots d\theta_N$. Next we note that, in these latter coordinates, the unit sphere in \mathbb{C}^N is defined by the condition $R = 1$. Thus the usual area element, dA , on our unit sphere is obtained by setting $R = 1$ and removing the term dR from this formula. i.e.

$$dA = r_1 \dots r_{N-1} dr_1 \dots dr_{N-1} d\theta_1 \dots d\theta_N.$$

It is now convenient to replace the coordinates r_a ($a = 1, \dots, N-1$) by w_a ($a = 1, \dots, N-1$) where $w_a = r_a^2$, whereupon clearly

$$dA = 2^{N-1} dw_1 \dots dw_{N-1} d\theta_1 \dots d\theta_N.$$

In view of the relation between the w_a and the first $N-1$ of the r_a and the fact that the r_a satisfy (88) with $R = 1$, the $(N-1)$ -tuple (w_1, \dots, w_{N-1}) clearly takes values which range over the simplex defined by the inequalities $0 \leq w_a$ for each a -value from 1 to $N-1$, together with the inequality $\sum_{a=1}^{N-1} w_a \leq 1$. (We shall call this the *standard $(N-1)$ -simplex*.) We remark that we can think of the quantity $1 - \sum_{a=1}^{N-1} w_a$ as ‘ w_N ’ for we will then have $w_N^2 = r_N^2$. So the latter inequality can then be expressed as $0 \leq w_N$. The θ_a values each, of course, range over $[0, 2\pi)$. So the area of our unit sphere is the integral of dA over the above ranges for our variables which is $2^{N-1}(2\pi)^N$ times the volume of our simplex. But the latter is easily seen to be $1/(N-1)!$. So the area of our unit sphere is $2\pi^N/(N-1)!$ which, of course, is the well-known value for the surface area of the real $(2N-1)$ -sphere. We want our Haar measure to be a probability measure, so we need to normalize it by dividing by this surface area. In conclusion, (up to an irrelevant set of measure zero) we have coordinatized the set of normalized vectors in our N -dimensional Hilbert space by products of $(N-1)$ -tuples (w_1, \dots, w_{N-1}) whose values range over our standard $(N-1)$ -simplex, with N -tuples $(\theta_1, \dots, \theta_N)$ whose values

range over the standard (i.e. with all periods equal to 2π) N -torus, and, with this coordinatization, (normalized) Haar measure is simply the product

$$d\text{Haar} = d(\text{Simplex}) \times d(\text{Torus}) \quad (89)$$

where

$$d\text{Simplex} = (N-1)!dw_1 \dots dw_{N-1} \quad (90)$$

and

$$d\text{Torus} = (2\pi)^{-N}d\theta_1 \dots d\theta_N. \quad (91)$$

For later convenience, we next define, and record, the easy-to-check values of, certain integrals of certain products of w 's over our simplex w.r.t. $d\text{Simplex}$:

$$J_1 := \int w_1 d\text{Simplex} = 1/N, \quad (92)$$

$$J_{11} := \int w_1^2 d\text{Simplex} = 2/N(N+1), \quad (93)$$

$$J_{12} := \int w_1 w_2 d\text{Simplex} = 1/N(N+1). \quad (94)$$

Obviously we assume here, for J_1 and J_{11} , that N is at least 2, and for J_{12} , that N is at least 3. We note that J_p (defined as J_1 but with w_1 replaced by w_p) will equal J_1 for any other value of p between 1 and N . Similarly (and with an obvious corresponding notation) $J_{pp} = J_{11}$ for any other p between 1 and N , and $J_{qp} = J_{12}$ for any pair of *different* q and p between 1 and N . We reiterate that all this holds even if q or p is equal to N , in which case, as we remarked above, w_N is taken to mean $1 - w_1 - \dots - w_{N-1}$.

We next use this coordinatization to compute the average, $\langle \rho_m^{mn} \rangle$, of ρ_m^{mn} (see the paragraph before Equation (83) in Section X) over Haar measure (with the result (97) below) and also to (re-)derive Lubkin's formula (83) for $\langle \text{tr}((\rho_m^{mn})^2) \rangle$. (Here, as in Section X, we indicate averages with respect to Haar measure with angle-brackets $\langle \ \rangle$.) Let Ψ be an arbitrary unit vector in $\mathcal{H}_m \otimes \mathcal{H}_n$ and choose (arbitrary) bases, $\{e_1, \dots, e_m\}$ for \mathcal{H}_m and $\{f_1, \dots, f_n\}$ for \mathcal{H}_n . Then we may write

$$\Psi = c_{ak} e_a \otimes f_k \quad (\text{summed over } a \text{ and } k). \quad (95)$$

(where $c_{ak} c_{ak}^*$ [summed over a and k] = 1) and the reduced density operator (see Sections I and X) $\check{\rho}_m^{mn}$ on \mathcal{H}_m takes the form $(\check{\rho}_m^{mn})_{a\hat{a}} |e_a\rangle \langle e_{\hat{a}}|$ (summed over a and \hat{a} from 1 to m) where

$$(\check{\rho}_m^{mn})_{a\hat{a}} = c_{ak} c_{\hat{a}k}^* \quad (\text{summed over } k). \quad (96)$$

(The reason for the 'check' ' $\check{\cdot}$ ' is that we will also want, below, to talk about the $(m \times m)$ matrix whose components are $(\rho_m^{mn})_{a\hat{a}}$. And we call this ' $\check{\rho}_m^{mn}$ ' to distinguish it from the operator ρ_m^{mn} on \mathcal{H}_m .) We want to average this over the unit sphere in \mathbb{C}^N for $N = mn$ where each factor of \mathbb{C} accommodates one of the $N = mn$ components of c_{ak} . So we replace c_{ak} in (96) by $r_{ak} e^{i\theta_{ak}}$ and then by $w_{ak}^{1/2} e^{i\theta_{ak}}$ and similarly for $c_{\hat{a}k}$, obtaining

$$(\check{\rho}_m^{mn})_{a\hat{a}} = w_{ak}^{1/2} w_{\hat{a}k}^{1/2} e^{i(\theta_{ak} - \theta_{\hat{a}k})} \quad (\text{summed over } k)$$

and we integrate this over Haar measure (89). Integrating over the θ s first (with $d\text{Torus}$ (91)) will obviously give a factor of $\delta_{a\hat{a}}$ for each k in the sum (from 1 to n) over k . We are thus left with a sum (over k) of n integrals,

$$\int w_{ak} d\text{Simplex}$$

for each a , each of which takes the form of J_1 (92) for $N = mn$. So we conclude that

$$\langle (\check{\rho}_m^{mn})_{a\hat{a}} \rangle = \frac{n\delta_{a\hat{a}}}{mn} = \frac{1}{m}\delta_{a\hat{a}}$$

and hence, obviously,

$$\langle \rho_m^{mn} \rangle = \frac{1}{m} I_m \quad (97)$$

where I_m denotes the identity operator on \mathcal{H}_m . Similarly, denoting, by ρ_n^{mn} the reduced density operator on \mathcal{H}_n we will have

$$\langle \rho_n^{mn} \rangle = \frac{1}{n} I_n. \quad (98)$$

We remark that we don't strictly need this result for the qualitative consequences of Lubkin's results we discuss below, but it is anyway interesting and also serves as a useful preliminary to the recalculation of Lubkin's result to which we will turn next – see especially the remark after equation (106). More importantly, we will need the counterpart to this result in our argument, below, for the closeness of ρ_S^{modern} and $\rho_S^{\text{modapprox}}$.

Proceeding similarly, it is straightforward to see that

$$\begin{aligned} \text{tr}((\rho_m^{mn})^2) &= c_{ak} c_{\hat{a}k}^* c_{al} c_{\hat{a}l}^* \quad (\text{summed over } k, l, a \text{ and } \hat{a}) \\ &= w_{ak}^{1/2} w_{\hat{a}k}^{1/2} w_{al}^{1/2} w_{\hat{a}l}^{1/2} e^{i(\theta_{ak} - \theta_{\hat{a}k})} e^{-i(\theta_{al} - \theta_{\hat{a}l})} \quad (\text{summed over } k, l, a \text{ and } \hat{a}). \end{aligned}$$

We integrate this over $d\text{Haar}$, again doing the θ -integrals first. Clearly the latter will vanish unless either $a = \hat{a}$ or $k = l$ (or both) whereupon, for fixed values of a, \hat{a}, k and l , the complex exponential will integrate (with $d\text{Torus}$ (91)) to 1. Moreover, (i) If $a = \hat{a}$ and $k = l$, then the w-integral over the simplex will equal J_{11} (93) for $N = mn$ – and there are mn such cases; (ii) If $a \neq \hat{a}$ and $k = l$, then the w-integral over the simplex will equal J_{12} (94) for $N = mn$ – and there are $nm(m-1)$ such cases; and finally (iii) If $a = \hat{a}$ and $k \neq l$, then the w-integral over the simplex will equal J_{12} for $N = mn$ again, and there are $n(n-1)m$ such cases. Thus we conclude that

$$\langle \text{tr}((\rho_m^{mn})^2) \rangle = \frac{2mn + mn(m-1) + n(n-1)m}{mn(mn+1)} = \frac{m+n}{mn+1} \quad (99)$$

in agreement with (83).

Lubkin's result (83)/(99) is important for us because of two qualitative consequences: First, as we mentioned in Section X and as Lubkin himself essentially argued, if $n \gg m$ then $\langle \text{tr}((\rho_m^{mn})^2) \rangle$ will be close to $1/m$. The only way this can happen is if *most* (in the sense we clarify below) totem states have reduced system density operators ρ_m^{mn} close to the maximally mixed density operator $(1/m)I_m$. To see this, notice that (adopting the convention of counting each eigenvalue, λ_a , ν times when ν is its multiplicity) amongst density operators, ρ , on an m -dimensional Hilbert space, the eigenvalues of ρ have to satisfy both $\sum_{a=1}^m \lambda_a = 1$ (since every density operator has unit trace) as well as $\sum_{a=1}^m \lambda_a^2 = \text{tr}(\rho^2)$ and one easily sees from these two conditions that the minimum value of $\text{tr}(\rho^2)$ is $1/m$ and that this minimum value is attained only when each of the λ_a equals $1/m$. If we next consider the set of such ρ for which $\text{tr}(\rho^2)$ is equal to $1/m + \eta$ where η denotes a (small) positive number, then one easily sees (again by considering the sum of the eigenvalues and the sum of their squares) that each of the λ_a must take the form $1/m + \delta_a$ where $\sum_{a=1}^m \delta_a^2 = \eta$. Applying this result to each of our reduced density operators ρ_m^{mn} , now writing the eigenvalues of each of these in the form $1/m + \delta_a$, then we immediately see that if $\langle \text{tr}((\rho_m^{mn})^2) \rangle = 1/m + \eta$ (which will hold with $\eta = (m+n)/(mn+1) - 1/m = (m^2-1)/(mn+1)$ which will be small if $n \gg m$) then the statement in words:

$$\text{For } n \gg m, \rho_m^{mn} \text{ will probably be close to } \frac{1}{m} I_m \quad (100)$$

will hold in the sense that $\sum_{a=1}^m \langle \delta_a^2 \rangle = \eta$.

Similar results will obviously hold for ρ_n^{mn} :

$$\text{For } m \gg n, \rho_n^{mn} \text{ will probably be close to } \frac{1}{n} I_n \quad (101)$$

in a similar sense to above.

Our second qualitative consequence of Lubkin's result for ρ_m^{mn} arises as a corollary to the above statement about ρ_m^{mn} : To explain what it is, note first that, just as we had the formula (96) for the components, $(\tilde{\rho}_m^{mn})_{a\hat{a}}$, of the $m \times m$ matrix $\tilde{\rho}_m^{mn}$, so we clearly have that $\rho_n^{mn} = (\tilde{\rho}_n^{mn})_{k\hat{k}} |f_k\rangle \langle f_{\hat{k}}|$ where, in the notation of (95), the $n \times n$ matrix $\tilde{\rho}_n^{mn}$ is given by

$$(\tilde{\rho}_n^{mn})_{k\hat{k}} = c_{ak} c_{\hat{a}k}^* \quad (\text{summed over } a). \quad (102)$$

So, denoting by C the $(m \times n)$ matrix whose components are the c_{ak} and by C^+ its $(n \times m)$ adjoint matrix, we clearly have

$$\tilde{\rho}_m^{mn} = CC^+ \text{ and } \rho_n^{mn*} = C^+C. \quad (103)$$

It easily follows from (103) that $x \in \mathbb{C}^m$ can be an eigenvector of $\tilde{\rho}_m^{mn}$ with a non-zero (positive) eigenvalue, λ , if and only if $y = \lambda^{-1/2}(C^+x)^*$ ($\in \mathbb{C}^n$) is an eigenvector of ρ_n^{mn} with the same eigenvalue. (The factor of $\lambda^{-1/2}$ is easily seen to be needed if we want to ensure that y is normalized whenever x is normalized.) Moreover we note for future reference (in our digression on the Schmidt decomposition below) that we then have $Cy^* = \lambda^{1/2}x$ – i.e.

$$c_{ak}y^{k*} = \lambda^{1/2}x^a \quad (104)$$

– the left hand side being summed over k . We conclude (continuing to adopt the convention of counting any eigenvalue ν times if it has multiplicity ν) that, if $m > n$ and if $\tilde{\rho}_m^{mn}$ has eigenvalues $\lambda_1, \dots, \lambda_n$, then $\tilde{\rho}_m^{mn}$ will have this same set of eigenvalues together with $m - n$ more, all of which will, however, be zero! Moreover (cf. the discussion after equation (100)), since $\langle \text{tr}(\rho_n^{mn})^2 \rangle = 1/n + \eta$ for η now equal to $(m+n)/(mn+1) - 1/n = (n^2 - 1)/(mn+1)$ – which will be small if $m \gg n$ – we will have $\lambda_k = 1/n + \delta_k$ where $\sum_{k=1}^n \langle \delta_k^2 \rangle = \eta$. So, we may say that

$$\text{For } m \gg n, \rho_m^{mn} \text{ will probably be close to } \frac{1}{n} \sum_{k=1}^n |\tilde{e}_a\rangle\langle\tilde{e}_a| \quad (105)$$

where $\{\tilde{e}_1, \dots, \tilde{e}_n\}$ is a basis for an n -dimensional subspace of \mathcal{H}_m (which will depend on Ψ). This is the second qualitative consequence of Lubkin's result we promised to arrive at at the outset. As far as we are aware, it does not appear to have been pointed out before. But, for our purposes, it will be of equal importance to the first consequence.

Similarly, of course:

$$\text{For } n \gg m, \rho_n^{mn} \text{ will probably be close to } \frac{1}{m} \sum_{a=1}^m |\tilde{f}_a\rangle\langle\tilde{f}_a| \quad (106)$$

where $\{\tilde{f}_1, \dots, \tilde{f}_m\}$ is a basis for an m -dimensional subspace of \mathcal{H}_n (which will again depend on Ψ).

We remark that, in preparation for the argument we give below for the claim that ρ_S^{modern} is well-approximated by $\rho_S^{\text{modapprox}}$, it is useful to observe that/how (105) and (101) are consistent with (97) and (98).

Further insight into the origin of (100), (101), (105) and (106) can be had by recalling that a given vector $\Psi \in \mathcal{H}_m \otimes \mathcal{H}_n$ – which we have written so far in the form (95) – can also be written as a single sum

$$\Psi = \sum_{i=1}^{\min(m,n)} \lambda_i^{1/2} \tilde{e}_i \otimes \tilde{f}_i \quad (107)$$

for suitable choices of basis $\{\tilde{e}_1, \dots, \tilde{e}_m\}$ on \mathcal{H}_m and $\{\tilde{f}_1, \dots, \tilde{f}_n\}$ on \mathcal{H}_n . This is the well-known Schmidt decomposition (cf. e.g. [49] and/or the next paragraph) and the λ_i are the same as those discussed above. (100) and (106) may then be viewed as (easy) consequences of the fact that, when $n \gg m$, the λ_i in (107) are probably close to $1/m$ for $i = 1, \dots, m$, while they are zero for $i > m$. Similarly (101) and (105) may be viewed as consequences of the fact that, when $m \gg n$, the λ_i in (107) are probably close to $1/n$ for $i = 1, \dots, n$, while they are zero for $i > n$.

The Schmidt decomposition in the form (107) can actually be derived easily from the results following Equation (103). In this paragraph, we digress to point out how, treating the cases where $n \geq m$: Denote by $\{x_1, \dots, x_m\}$ a complete set of orthonormal eigenvectors of $\tilde{\rho}_m^{mn}$, and denote by x_i^a the a th component of x_i . In view of the sentence following Equation (103), we can clearly find a complete set, $\{y_1^*, \dots, y_n^*\}$, of orthonormal eigenvectors of ρ_n^{mn} so that, denoting by y_j^{k*} the k th component of y_j^* , we have, by (104), $c_{ak}y_j^{k*} = \lambda_j^{1/2}x_j^a$ (the left hand side being summed over k). (We only need to make sure that the i -value of every x_i belonging to each non-zero eigenvalue of $\tilde{\rho}_m^{mn}$ matches the j -value of a y_j^* belonging to an equal non-zero eigenvalue of ρ_n^{mn} – there being necessarily an equal number of each; for any other y_j [and of course there have to be others whenever $n > m$] the right hand side will anyway vanish.) Also introduce a new basis $\{\tilde{e}_1, \dots, \tilde{e}_m\}$ for \mathcal{H}_m such that $e_a = x_i^a \tilde{e}_i$ (summed over i) and, similarly, introduce a new basis $\{\tilde{f}_1, \dots, \tilde{f}_n\}$ for \mathcal{H}_n such that $f_k = y_j^{k*} \tilde{f}_j$ (summed over j). Then we have (see (95))

$$\Psi = c_{ak}e_a \otimes f_k \text{ (summed over } a \text{ and } k) = c_{ak}x_i^a y_j^{k*} \tilde{e}_i \otimes \tilde{f}_j \text{ (summed over } i, j, a \text{ and } k)$$

$$= \text{by (104)} \lambda_j^{1/2} x_i^a x_j^a \tilde{e}_i \otimes \tilde{f}_j \text{ (summed over } i, j \text{ and } a) = \lambda_j^{1/2} \delta_{ij} \tilde{e}_i \otimes \tilde{f}_j \text{ (summed over } i \text{ and } j)$$

So we have

$$\Psi = \lambda_i^{1/2} \tilde{e}_i \otimes \tilde{f}_i \text{ (summed over } i \text{)}$$

thus establishing (107) in cases where $n \geq m$. The cases where $m \geq n$ are obviously similar. This ends our digression.

XII. MAIN ARGUMENT FOR THE VALIDITY OF THE APPROXIMATION (15) OF ρ_S^{modern} BY $\rho_S^{\text{modapprox}}$

Let us now turn to consider the set of unit vectors, Ψ , in the Hilbert space of a totem as specified in Section I C – i.e. in the subspace of states with total energies in the range $[E, E + \Delta]$ of $\mathcal{H}_S \otimes \mathcal{H}_B$. Allowing ourselves to make the slight distortion explained before equations (4) and (5), we may assume \mathcal{H}_S has an orthonormal basis consisting of vectors $|\epsilon_S, i\rangle$, where ϵ_S ranges from Δ to E in steps of Δ , while, for each ϵ_S , the integer, i , ranges from 1 to $n_S(\epsilon_S)$; and similarly \mathcal{H}_B has an orthonormal basis consisting of vectors $|\epsilon_B, j\rangle$, where ϵ_B ranges from Δ to E in steps of Δ , while, for each ϵ_B , j ranges from 1 to $n_B(\epsilon_B)$. (Below, and as in Section I, we shall sometimes drop the S and B subscripts on the ϵ s when no ambiguity can arise.) Then, we are interested in the set of unit vectors, Ψ , in the subspace, which we shall call below \mathcal{H}_M , of $\mathcal{H}_S \otimes \mathcal{H}_B$ with total energy exactly E . The reason for the name \mathcal{H}_M is that \mathcal{H}_M will clearly have dimension M , where M is as in (7).

Each such $\Psi \in \mathcal{H}_M$ is writeable in the form

$$\Psi = \sum_{\epsilon=\Delta}^E \sum_{i=1}^{n_S(\epsilon)} \sum_{j=1}^{n_B(E-\epsilon)} c_{\epsilon}^{ij} |\epsilon, i\rangle \otimes |E - \epsilon, j\rangle \quad (108)$$

where we recall (see above and cf. before Equation (6)) that the sum over ϵ goes up in integer multiples of Δ . We also note that since Ψ is a unit vector, the sum (with the above indicated ranges) over ϵ , i and j of $|c_{\epsilon}^{ij}|^2$ equals 1. For such a Ψ , the partial trace of $|\Psi\rangle\langle\Psi|$ over \mathcal{H}_B , i.e. the reduced density operator, ρ_S^{modern} , on \mathcal{H}_S , will then clearly be given by

$$\rho_S^{\text{modern}} = \sum_{\epsilon=\Delta}^E \sum_{i=1}^{n_S(\epsilon)} \sum_{\hat{i}=1}^{n_S(\epsilon)} \tilde{r}_{\epsilon}^{S^{i\hat{i}}} |\epsilon, i\rangle\langle\epsilon, \hat{i}|, \quad (109)$$

where

$$\tilde{r}_{\epsilon}^{S^{i\hat{i}}} = \sum_{j=1}^{n_B(E-\epsilon)} c_{\epsilon}^{ij} c_{\epsilon}^{\hat{i}j*}. \quad (110)$$

We shall find it useful sometimes to think of \mathcal{H}_S as a direct sum $\oplus_{\epsilon=0}^E \mathcal{H}_{\epsilon}^S$ (and similarly for \mathcal{H}_B) where \mathcal{H}_{ϵ}^S is spanned by the $|\epsilon, i\rangle$ for fixed ϵ as i varies from 1 to $n_S(\epsilon)$ and we shall call the restriction of ρ_S^{modern} to \mathcal{H}_{ϵ}^S simply r_{ϵ}^S – its $i\hat{i}$ components in the basis consisting of the $|\epsilon, i\rangle$ being obviously the $\tilde{r}_{\epsilon}^{S^{i\hat{i}}}$ introduced above.

Our aim is to give an argument in favour of the claimed correctness of the proposition, which we state in Section I D, that, in situations of interest, ρ_S^{modern} will be well-approximated by the $\rho_S^{\text{modapprox}}$ of (15) and, in the course of giving this argument, to make clear how the $n_B(E - \epsilon)$ -dimensional subspace of the $(n_S(\epsilon)$ -dimensional) energy- ϵ subspace of \mathcal{H}_S spanned by the $|\epsilon, i\rangle$ depend on Ψ . (We should amplify on this statement by explaining that, when we say that ρ_S^{modern} is well-approximated by $\rho_S^{\text{modapprox}}$, what we mean is that the values of physical quantities of interest, such as the mean energy and the entropy of the system S, calculated using ρ_S^{modern} will be close to the values of the same quantities calculated using $\rho_S^{\text{modapprox}}$.)

The main ingredients in our argument concern the average, $\langle r_{\epsilon}^S \rangle$, of r_{ϵ}^S and also the average, $\langle \text{tr}((r_{\epsilon}^S)^2) \rangle$, of $\text{tr}((r_{\epsilon}^S)^2)$, where both averages are taken as Ψ ranges over the whole of \mathcal{H}_M (with respect to Haar measure on \mathcal{H}_M). The calculations of these quantities are closely similar to the preliminary calculations we carried out above for the average of the density operator ρ_m^{mn} and the average of the trace of its square; the difference being that we are now averaging over all unit Ψ in our full M -dimensional Hilbert space \mathcal{H}_M (with M as in (7)) even though what we are averaging is only the restriction, r_{ϵ}^S , (and the trace of the square of the restriction) of ρ_S^{modern} to \mathcal{H}_{ϵ}^S for fixed ϵ . As a result, while the counterpart of the product, mn , in the denominator in our preliminary calculation would just be the single product $n_S(\epsilon)n_B(E - \epsilon)$ were our average only to be over $\mathcal{H}_{\epsilon}^S \otimes \mathcal{H}_{E-\epsilon}^B$, since we average over the unit vectors of the full Hilbert space \mathcal{H}_M , the counterpart will be turn out to be M . Aside from this difference, to calculate $\langle r_{\epsilon}^S \rangle$ one proceeds

very similarly to the passage, above, between equations (96) and (97); the reader can easily supply the details simply by replacing (96) by (110) etc. One clearly obtains (instead of (97))

$$\langle r_\epsilon^S \rangle = \frac{n_B(E - \epsilon)}{M} I_\epsilon^S \quad (111)$$

where I_ϵ^S denotes the identity on \mathcal{H}_ϵ^S . Similarly, proceeding as in the passage between equations (97) and (99) (but again it will turn out that one needs to replace mn in the denominator by M) we easily find:

$$\begin{aligned} \langle \text{tr}((r_\epsilon^S)^2) \rangle &= \frac{2n_B(E - \epsilon)n_S(\epsilon) + n_B(E - \epsilon)n_S(\epsilon)(n_S(\epsilon) - 1) + n_B(E - \epsilon)(n_B(E - \epsilon) - 1)n_S(\epsilon)}{M(M + 1)} \\ &= \frac{n_B(E - \epsilon)n_S(\epsilon)(n_B(E - \epsilon) + n_S(\epsilon))}{M(M + 1)}. \end{aligned} \quad (112)$$

Now, rather as in our arguments for the two qualitative consequences of Lubkin's result, (but now our arguments will involve both the counterpart, (111), to (97) as well as the counterpart, (112), to (99)) we observe from (112) that, whenever $n_S(\epsilon) \gg n_B(E - \epsilon)$, $\langle \text{tr}((r_\epsilon^S)^2) \rangle$ will be very close to $M^{-2}(n_B(E - \epsilon))^2$, which, *in the presence of* (111), easily implies (i.e. by similar reasoning to that used above in our derivation of (100) and (101)) that r_ϵ^S must be very close to $M^{-1}n_B(E - \epsilon) \sum_{i=1}^{n_S(\epsilon)} |\epsilon, i\rangle \langle \epsilon, i|$. Moreover, whenever $n_B(E - \epsilon) \gg n_S(\epsilon)$, $\langle \text{tr}((r_\epsilon^S)^2) \rangle$ will be very close to $M^{-2}(n_S(\epsilon))^2$, which, again in the presence of (111), easily implies (i.e. by similar reasoning to that used above in our derivation of (105) and (106)) that r_ϵ^S will be very close to $M^{-1}n_S(\epsilon) \sum_{i=1}^{n_B(E - \epsilon)} |\widetilde{\epsilon}, i\rangle \langle \widetilde{\epsilon}, i|$ where $|\widetilde{\epsilon}, i\rangle$ denote the elements of an orthonormal basis for an $n_B(E - \epsilon)$ -dimensional subspace of the $(n_S(\epsilon))$ -dimensional energy- ϵ subspace, \mathcal{H}_ϵ^S of \mathcal{H}_S which will depend on Ψ . Comparing these conclusions with the form of Equation (15) we immediately see that, if it were the case that for all ϵ , either $n_S(\epsilon) \gg n_B(E - \epsilon)$ or $n_B(E - \epsilon) \gg n_S(\epsilon)$, then (15) would obviously be a good approximation (at least each term in the sum over ϵ will be) for all ϵ . However, of course, in typical situations of interest, there will be a region of ϵ values around the value E_c – see the definitions of terms immediately after Equation (15) – where neither of these statements will hold and $n_S(\epsilon)$ and $n_B(E - \epsilon)$ will be of comparable size. (We remark in passing, though, that, typically, [one will be able to choose Δ so that] each of these quantities will be much greater than 1 for all or very nearly all ϵ [which are multiples of Δ and] in the range $[0, E]$.)

Nevertheless for the sort of situations of interest to us – and, in particular, for the densities of states which increase according to the power law, (18), as considered in Section II or which increase exponentially, (34), as considered in Sections III and V, or which increase as quadratic exponentials, (58), as considered in Section VI – and assuming the totem energy E and our choice of energy-increment, Δ (see Section I) are such that $M \gg 1$ (to ensure that the system [and bath] has access to a very large number of states) – one can check that the region of ϵ -values around E_c where $n_B(E - \epsilon)$ and $n_S(\epsilon)$ are of comparable size will always be very small in size compared to E , while the sum over this region of $n_S(\epsilon)n_B(E - \epsilon)$ will be very small compared to M . (In other words, the integral over this energy-region of the energy-probability density $P_S(\epsilon)$ [see (10)] will be very much less than 1.) Moreover, as ϵ decreases towards zero, or increases towards E from E_c , then for all three densities of states, (18), (34), (58), one may check that the ratio $n_S(\epsilon)/n_B(E - \epsilon)$, respectively $n_B(E - \epsilon)/n_S(\epsilon)$, and hence the counterparts (i.e. with $n_S(\epsilon)$ replacing m and $n_B(E - \epsilon)$ replacing n) to the quantities which we called η before Equation (100), respectively Equation (105), will get rapidly smaller and hence the relevant notion of closeness (i.e. as in (100), (105)) will get rapidly stronger. It is then straightforward to argue from these statements that quantities of interest such as (cf. (38)) $\bar{\epsilon}_S^{\text{modern}} =: \text{tr}(\rho_S^{\text{modern}} H_S)$, $\text{tr}((\rho_S^{\text{modern}})^2)$ itself, and (cf. (41)) $S_S^{\text{modern}} =: -\text{tr}(\rho_S^{\text{modern}} \log \rho_S^{\text{modern}})$ will be closely approximated (respectively) by $\bar{\epsilon}_S^{\text{modapprox}} =: \text{tr}(\rho_S^{\text{modapprox}} H_S)$, $\text{tr}((\rho_S^{\text{modapprox}})^2)$, and (cf. (41)) $S_S^{\text{modapprox}} =: -\text{tr}(\rho_S^{\text{modapprox}} \log \rho_S^{\text{modapprox}})$.

Concerning the latter two quantities – i.e. the trace of the square of the reduced density operator of S and its von Neumann entropy – there are reasons to expect the approximation of S_S^{modern} by $S_S^{\text{modapprox}}$ to be even better than the approximation of $\text{tr}((\rho_S^{\text{modern}})^2)$ by $\text{tr}((\rho_S^{\text{modapprox}})^2)$.

XIII. FURTHER CHECKS AND DETAILS ON THE VALIDITY OF THE APPROXIMATION (15)

As a partial check of various aspects of all of the above argument, and in justification of our latter remark, it is instructive first to consider the case where, for all ϵ ($= 0, \Delta, 2\Delta, \dots$) in the range $[0, E]$, we have $n_S(\epsilon) = 1 = n_B(E - \epsilon)$ where, of course it is *never* true that $n_S(\epsilon) \gg n_B(E - \epsilon)$ or that $n_S(\epsilon) \ll n_B(E - \epsilon)$ (nor that each of these quantities is very much greater than 1!) so we can think of this as one sort of 'worst case scenario'. Of course this is not an example that interests us in Part 1, but it would apply e.g. to a totem consisting of a pair of weakly coupled quantum

harmonic oscillators (with equal spring constants) for a total energy much greater than the level spacing (and a choice of Δ equal to the level spacing). For this model, we may clearly write

$$\Psi = \sum_{\epsilon=\Delta}^E c_\epsilon |\epsilon\rangle |E - \epsilon\rangle \quad (113)$$

so that the reduced density operator of the system, S, will be

$$\rho_S^{\text{modern}} = \sum_{\epsilon=\Delta}^E |c_\epsilon|^2 |\epsilon\rangle \langle \epsilon|,$$

while the formula, (15), for $\rho_S^{\text{modapprox}}$ (15) becomes simply

$$\rho_S^{\text{modapprox}} = \frac{1}{M} \sum_{\epsilon=\Delta}^E |\epsilon\rangle \langle \epsilon|,$$

where (cf. (7)) $M = E/\Delta$. Clearly, (cf. (38)) the approximate mean energy,

$$\begin{aligned} \bar{\epsilon}_S^{\text{modapprox}} &:= \text{tr}(\rho_S^{\text{modapprox}} H_S) = \frac{1}{M} \sum_{\epsilon=\Delta}^E \epsilon \\ &= \frac{1}{2M} E \left(\frac{E}{\Delta} + 1 \right) = \frac{E}{2}. \end{aligned}$$

(In calculating the value of the above sum, we need of course to recall that the sum is over values of ϵ which are integer multiples of Δ .) Therefore, since this doesn't depend on the c_ϵ , its average over Haar measure (indicated with “ $\langle \rangle$ ”) takes the same value:

$$\langle \bar{\epsilon}_S^{\text{modapprox}} \rangle = \frac{E}{2}.$$

On the other hand, the average over Haar measure of the exact mean energy, $\bar{\epsilon}_S^{\text{modern}}$, may be calculated as follows:

$$\begin{aligned} \langle \bar{\epsilon}_S^{\text{modern}} \rangle &= \left\langle \sum_{\epsilon=\Delta}^E \epsilon |c_\epsilon|^2 \right\rangle \\ &= \sum_{\epsilon=\Delta}^E \epsilon \int w_\epsilon d\text{Haar} \end{aligned}$$

where the integral is over the complex M -dimensional sphere of unit vectors in the Hilbert space, \mathcal{H}_M , spanned by the vectors of form (113), coordinatized with w_ϵ ranging over the $(M-1)$ -simplex and θ_ϵ ranging over the M -torus as explained at the beginning of Section XI, where $w_\epsilon = |c_\epsilon|^2$ etc. Obviously the torus factor of the integral just gives 1, so the integral has, by (92), the value $1/M$ for each ϵ . So we conclude that $\langle \bar{\epsilon}_S^{\text{modern}} \rangle$ has the same value as $\langle \bar{\epsilon}_S^{\text{modapprox}} \rangle$, i.e.

$$\langle \bar{\epsilon}_S^{\text{modern}} \rangle = \frac{E}{2},$$

which, of course, has to be the correct value by the symmetry under the interchange of S and B in this case.

Turning to averages over Haar measure of the trace of the square of the reduced density operator of S, we have, on the one hand,

$$\langle \text{tr}((\rho_S^{\text{modapprox}})^2) \rangle = \left\langle \sum_{\epsilon=\Delta}^E \frac{1}{M^2} \right\rangle = \sum_{\epsilon=\Delta}^E \frac{1}{M^2} = \left(\frac{\Delta}{E} \right)^2 \left(\frac{E}{\Delta} \right)$$

$$= \frac{1}{M} \quad (= \frac{\Delta}{E}).$$

Whereas, on the other hand,

$$\langle \text{tr}((\rho_S^{\text{modern}})^2) \rangle = \langle \sum_{\epsilon=\Delta}^E |c_\epsilon|^4 \rangle = \left(\frac{E}{\Delta} \right) \int w_1^2 d\text{Haar} = M \frac{2}{M(M+1)} \simeq \frac{2}{M} \quad (\simeq \frac{2\Delta}{E})$$

(where we have used (93) in calculating the integral) which differs from the approximate value by a factor of 2!

However, if we turn to calculate the averages over Haar measure of the von Neumann entropies of the approximate and exact reduced density operator of S, we find, on the one hand,

$$\langle S_S^{\text{modapprox}} \rangle = \langle -\text{tr}(\rho_S^{\text{modapprox}} \log \rho_S^{\text{modapprox}}) \rangle = \langle \log M \rangle = \log M \quad (= \log(E/\Delta)). \quad (114)$$

On the other hand,

$$\begin{aligned} \langle S_S^{\text{modern}} \rangle &= \langle -\text{tr}(\rho_S^{\text{modern}} \log \rho_S^{\text{modern}}) \rangle = -M \int_{\text{Unit Sphere in } \mathbb{C}^M} w_1 \log w_1 d\text{Haar} \\ &= \text{(by (89) and (90))} \quad -M(M-1)! \int_{\text{Simplex}} w_1 \log w_1 dw_1 \dots dw_{M-1} \\ &= -M(M-1) \int_0^1 w \log w (1-w)^{M-2} dw. \end{aligned}$$

One may do this integral by noticing that $w \log w = (dw^\alpha/d\alpha)|_{\alpha=1}$ – obtaining for its value, $d(B(\alpha+1, M-1))/d\alpha|_{\alpha=1}$ where B denotes the beta function (see e.g. [30]). One finds that $-M(M-1)$ times this simplifies (using (21)) to $\psi(1+M) - \psi(2)$ where ψ denotes the ψ (or ‘digamma’) function defined by $\psi(x) = d \log \Gamma(x)/dx$, and this [30], in turn, equals $\sum_{k=2}^M 1/k$ which, by the standard asymptotic expansion of Euler’s constant, C ($= 0.5772\dots$) is equal to $\log M + C - 1 + 1/2M + O(1/M^2)$. So we conclude that

$$\langle S_S^{\text{modern}} \rangle = \log M + C - 1 + O(1/M) \quad (= \log(E/\Delta) + C - 1 + O(\Delta/E)) \quad (115)$$

where C is Euler’s constant ($0.5772\dots$)

Comparison of (115) and (114) shows that the use of (15) for this ‘worst case scenario’ leads to a von Neumann entropy which, for large M , is very close to the average over Haar measure of its actual value. In view of the fact that both of these values are very close to the maximum possible value, $\log M$, of the entropy of any density operator on \mathcal{H}_S (which is of course M -dimensional in this case) we conclude both that most totem states, Ψ , for this model must have a reduced density operator on S whose von Neumann entropy is close to $\log M$; and that the use of (15) leads to a good approximation for this value. And both of these things hold even though, as we saw above, our general arguments do not apply to this case and even though, for this case, as we saw above, (15) leads to a trace of the square of the reduced density operator of S which is only half of the average over Haar measure of its actual value.

We will next use the Lubkin-Page asymptotic formula, (87), to obtain a result which tends to confirm the accuracy of our general formula (41), obtained using (15), for the von Neumann entropy for our densities of states of interest, (18), (34), (58). Our result will show that the value of the von Neumann entropy obtained with (15) well-approximates a certain restricted average of the exact von Neumann entropy. Before we present this result, we shall find it helpful to first explain what we mean here by a ‘restricted average’ in a different context:

Let us look back at the result essentially due to Lubkin, (99), which we (re-)obtained above, for the average over Haar measure on vectors, Ψ , belonging to the tensor product, $\mathcal{H}_m \otimes \mathcal{H}_n$, of two Hilbert spaces, of the trace of the square of the reduced density operator, ρ_m^{mn} on \mathcal{H}_m . Averaging over all totem vectors, $\Psi \in \mathcal{H}_m \otimes \mathcal{H}_n$, amounts, as we explained above, to averaging with the invariant measure on the complex mn -sphere over the coefficients, c_{ak} , in the basis-expansion, (95), of Ψ , which, in turn, writing c_{ak} as $w_{ak}^{1/2} e^{i\theta_{ak}}$, we saw, amounts to integrating w.r.t. the w_{ak} over the $(mn-1)$ -simplex and w.r.t. the θ_{ak} over the mn -torus. What we now wish to point out is that, if we restrict to c_{ak} which take the form $(1/\sqrt{mn})e^{i\theta_{ak}}$ and just average over these (i.e. by integrating with respect to the θ_{ak} over the mn -torus) one easily finds – denoting our restricted average with the symbol $[\]$ – that

$$[\text{tr}((\rho_m^{mn})^2)] = \frac{m+n-1}{mn},$$

and this is not a bad approximation to the value, $(m+n)/(mn+1)$, of the full average, $\langle \text{tr}((\rho_m^{mn})^2) \rangle$, of $\text{tr}((\rho_m^{mn})^2)$ – the two expressions differing, in fact, only by terms of order $1/mn$. In terms of our geometrical picture, in which averaging w.r.t. Haar measure amounts to integrating over the $(mn-1)$ -simplex times the mn -torus, in passing from the unrestricted average, $\langle \cdot \rangle$, to our restricted average, $[\cdot]$, we have replaced the integral over the $(mn-1)$ -simplex by the value at its centroid (where $w_{ak} = 1/mn$ for all a and k), but continue to integrate over the mn -torus. Of course this restricted average, $[\cdot]$, is a basis-dependent notion, but what we have learnt is that we can choose any bases $\{e_1, \dots, e_m\}$ and $\{f_1, \dots, f_n\}$ we like and we obtain this reasonably good approximation to the unrestricted average (at least for the quantity $\text{tr}((\rho_m^{mn})^2)$).

We conclude that the corresponding restricted set of totem vectors (cf. (95),

$$\Psi = \frac{1}{\sqrt{mn}} e^{i\theta_{ak}} e_a \otimes f_k \quad (\text{summed over } a \text{ and } k)$$

is (for any choice of bases, $\{e_1, \dots, e_m\}$ and $\{f_1, \dots, f_n\}$ and as the θ_{ak} range over the mn -torus) a sufficiently representative set of totem vectors for the restricted average over this set to indicate sufficiently well the behaviour of a generic totem state, Ψ (at least as far as $\text{tr}((\rho_m^{mn})^2)$ is concerned).

We shall proceed in a similar spirit, but now for the totem of Section IC. We expect that it won't make too big a difference if, instead of averaging over the full set of totem vectors, $\Psi \in \mathcal{H}_M$, we consider a suitable restricted average. To motivate the restriction that we shall make, we notice first, that, if we expand such vectors, Ψ , as in (108), then it follows from (111) that the (unrestricted) average value (i.e. over Haar measure on the set of all $\Psi \in \mathcal{H}_M$) of the trace of each r_ϵ^S (defined in the paragraph after (110)) is given by

$$\langle \text{tr}(r_\epsilon^S) \rangle = \left\langle \sum_{i=1}^{n_S(\epsilon)} \sum_{j=1}^{n_B(E-\epsilon)} |c_\epsilon^{ij}|^2 \right\rangle = \frac{n_S(\epsilon)n_B(E-\epsilon)}{M} \quad (= P_S(\epsilon)\Delta) \quad (116)$$

– the equality in parenthesis following from (110).

In view of this, we take our restricted average to be over vectors, $\Psi \in \mathcal{H}_M$, such that, in the expansion, (108), for each ϵ , the coefficients c_ϵ^{ij} are constrained to satisfy exactly

$$\sum_{i=1}^{n_S(\epsilon)} \sum_{j=1}^{n_B(E-\epsilon)} |c_\epsilon^{ij}|^2 \quad (= \text{tr}(r_\epsilon^S)) = \frac{n_S(\epsilon)n_B(E-\epsilon)}{M}.$$

(We remark that, in view of what we explained in the previous two paragraphs, we could alternatively restrict much further and simply average over Ψ in (108) for which every c_ϵ^{ij} takes the form $e^{i\theta_\epsilon^{ij}}/\sqrt{M}$ and still be able to arrive at similar conclusions to those below. However the restriction we adopt has the advantage of allowing us to directly use the Lubkin-Page approximation in exactly the form (87).) In other words, denoting $n_S(\epsilon)n_B(E-\epsilon)/M$ by μ_ϵ , we average over $\Psi \in \mathcal{H}_M$ which take the form $\oplus_{\epsilon=0}^E \sqrt{\mu_\epsilon} \Psi_\epsilon^M$ (each Ψ_ϵ^M being normalized) where we regard \mathcal{H}_M as the direct sum, $\oplus_{\epsilon=0}^E \mathcal{H}_\epsilon^M$, where, for each ϵ ($= 0, \Delta, \dots, E$), \mathcal{H}_ϵ^M denotes the $(n_S(\epsilon)n_B(E-\epsilon))$ -dimensional Hilbert subspace of \mathcal{H}_M spanned by the vectors $|\epsilon, i\rangle|E-\epsilon, j\rangle$, $i = 1 \dots n_S(\epsilon)$, $j = 1 \dots n_B(E-\epsilon)$ in $\mathcal{H}_\epsilon^S \otimes \mathcal{H}_{E-\epsilon}^B$ – see after equation (110). For such restricted Ψ , ρ_S^{modern} will take the form

$$\rho_S^{\text{modern}} = \sum_{\epsilon=\Delta}^E \mu_\epsilon R_\epsilon^S$$

where R_ϵ^S is the partial trace of $|\Psi_\epsilon^M\rangle\langle\Psi_\epsilon^M|$ over $\mathcal{H}_{E-\epsilon}^B$ (which will equal r_ϵ^S divided by its trace, which is μ_ϵ). Clearly, by the lemma in Section IV, we therefore have

$$S(\rho_S^{\text{modern}}) = S\left(\sum_{\epsilon=\Delta}^E \mu_\epsilon R_\epsilon^S\right) = \sum_{\epsilon=\Delta}^E \mu_\epsilon S(R_\epsilon^S) - \sum_{\epsilon=\Delta}^E \mu_\epsilon \log \mu_\epsilon. \quad (117)$$

But now we notice that, if we identify m with $n_S(\epsilon)$ and n with $n_B(E-\epsilon)$, then we can identify \mathcal{H}_ϵ^M with the Hilbert space, $\mathcal{H}_m \otimes \mathcal{H}_n$, of Sections I and XI, and, under this identification, R_ϵ^S is identified with ρ_m^{mn} , and $S(R_\epsilon^S)$ with $S(\rho_m^{mn})$. Moreover, averaging $S(R_\epsilon^S)$ over \mathcal{H}_ϵ^M is, under (the reverse of) this identification, then obviously the same as taking the unrestricted average of $S(\rho_m^{mn})$ over Haar measure on unit vectors in $\mathcal{H}_m \otimes \mathcal{H}_n$ and so we may estimate its value using the Lubkin-Page approximation (87). Making these identifications, if we now use $[\cdot]$ to denote

our restricted average over our restricted totem vectors, Ψ , and $\langle \cdot \rangle$ to denote the unrestricted average over Haar measure on unit vectors in \mathcal{H}_ϵ^M for each ϵ , we may calculate using the formula (117) in our lemma of Section IV:

$$[S(\rho_S^{\text{modern}})] = \left[S \left(\sum_{\epsilon=\Delta}^E \mu_\epsilon S(R_\epsilon^S) \right) \right] = \sum_{\epsilon=\Delta}^E \mu_\epsilon \langle S(R_\epsilon^S) \rangle - \sum_{\epsilon=\Delta}^E \mu_\epsilon \log \mu_\epsilon$$

which, recalling that $\mu_\epsilon = n_S(\epsilon)n_B(E-\epsilon)/M$ and using (87), equals

$$\begin{aligned} & \sum_{\epsilon=\Delta}^E \frac{n_S(\epsilon)n_B(E-\epsilon)}{M} \left(\log(\min(n_S(\epsilon), n_B(E-\epsilon))) - \frac{\min(n_S(\epsilon), n_B(E-\epsilon))}{2\max(n_S(\epsilon), n_B(E-\epsilon))} \right. \\ & \left. - \log \left(\frac{n_S(\epsilon)n_B(E-\epsilon)}{M} \right) + O \left(\frac{1}{n_S(\epsilon)n_B(E-\epsilon)} \right) \right) \end{aligned}$$

which easily simplifies to $[S(\rho_S^{\text{modern}})]$

$$\begin{aligned} & = -M^{-1} \left(\sum_{\epsilon=\Delta}^{E_c} n_S(\epsilon)n_B(E-\epsilon) \log(M^{-1}n_B(E-\epsilon)) + \sum_{\epsilon=E_c+\Delta}^E n_S(\epsilon)n_B(E-\epsilon) \log(M^{-1}n_S(\epsilon)) \right) \\ & \quad - M^{-1} \left(\sum_{\epsilon=\Delta}^{E_c} \frac{n_S(\epsilon)^2}{2} + \sum_{\epsilon=E_c+\Delta}^E \frac{n_B(E-\epsilon)^2}{2} \right) + O(1) \end{aligned} \quad (118)$$

where E_c is as defined after (15).

Comparing (118) with (41), we notice that the first line of (118) coincides with the formula, (41), $S_S^{\text{modapprox}}$ for the von Neumann entropy of $\rho_S^{\text{modapprox}}$ which we derived from (15). Thus we may conclude that our restricted average over totem vectors of S_S^{modern} will be given by the formula we gave for $S_S^{\text{modapprox}}$ in (41) – plus an ‘error term’ given by the last line of (118). Moreover, the close agreement found above between S_S^{modern} and $S_S^{\text{modapprox}}$ in the ‘worst case scenario’ discussed above, strongly suggests that the same statement will be true for the unrestricted average. In order to conclude that this amounts to an independent check of the correctness of the approximate formula S_S^{modern} of (41) for our densities of states of interest, (18), (34), (58), it remains to show that (./investigate when) the ‘error term’ (i.e. the second line in (118)) is small. To end this section we turn to this last question:

It is in fact easy to see (after converting the sum to an integral, using (8)) that: (a) for our power-law densities of states, (18), with $A_S = A_B$ and $N_S = N_B = N$ say, the last line of (118) (minus the $O(1)$ term) is (using Stirling’s approximation – see Section II) $1/\sqrt{\pi N}$; (b) for our (equal) exponential densities of states, (34), it is $(1/bE)(1 - e^{-bE})$; and (c) for our (equal) quadratic densities of states, (58), it is (approximating the integral with the leading term of the asymptotic formula in Endnote [42]) $\exp(-qE^2/2)$. These terms will all be much smaller than typical values of the first line of (118) provided N is large in (a), provided $E \gg 1/b$ (cf. Equation (36)) in (b), and provided $E \gg 1/\sqrt{q}$ (cf. Equation (59)) in (c).

So in all cases of interest here, and, no doubt, in many others too, the last line of (118) will be negligibly small.

Acknowledgments

I thank Michael Kay for useful discussions about traditional statistical mechanics and also for information about string theory and Richard Hall for showing me how to do integrals involving logarithms.

-
- [1] In the case the energy probability density is sharply peaked, $T(\epsilon)$ would then be interpretable as the temperature of a small subsystem or of a small system placed in thermal contact with the given system. See Section VII A.
 - [2] S.W. Hawking, *Commun. Math. Phys.* **43**, 199 (1975)
 - [3] S.W. Hawking, *Hawking on the Big Bang and Black Holes* (Advanced Series in Astrophysics and Cosmology 8) (World Scientific, Singapore 1993)
 - [4] G.W. Gibbons and S.W. Hawking (eds.), *Euclidean Quantum Gravity* (World Scientific, Singapore 1993)
 - [5] G.W. Ford, M. Kac and P. Mazur, *J. Math. Phys.* **6**, 504 (1965)

- [6] E.B. Davies, *Quantum Theory of Open Systems* (Academic, London 1976)
- [7] R.P. Feynman, *Statistical Mechanics* (W.A. Benjamin Inc., Reading Mass. 1972)
- [8] W. Thirring, *A Course in Mathematical Physics 4: Quantum Mechanics of Large Systems* (Springer, New York 1980)
- [9] S. Goldstein, J.L. Lebowitz, R. Tumulka and N. Zanghi, Phys. Rev. Lett. **96**, 050403 (2006) [arXiv:cond-mat/0511091]
- [10] E. Lubkin, J. Math. Phys. **19**, 1028 (1978)
- [11] Actually Lubkin was by no means the first to adopt the natural invariant measure ('Haar measure') in a related context. See the discussion and references in [9].
- [12] S. Popescu, A.J. Short and A. Winter, Nature Physics **2**, 754 (2006) [arXiv:quant-ph/0511225]
- [13] Throughout the paper, we will refer to explanations of the origin of thermality, based on the assumption of a total pure state, as being 'modern' (in contradistinction to the 'traditional' assumption of a microcanonical ensemble).
- [14] Throughout this paper, we shall often consider, side-by-side with our finite-sum formulae, continuum approximations where we make the replacement (8). Usually, the formulae are essentially interchangeable and one can (and it can be helpful to) check they give the same answers. However, we caution that, for example, the continuum versions of the entropy formulae which we obtain, using the prescription (8), in Section IE can end up having energy-intervals on which the 'number of energy levels' is greater than zero but less than 1 and this can result in spurious negative terms which go like the logarithm of (a constant with the dimensions of inverse energy times the) total energy E . An example of such a term occurs in Section VI as is mentioned there.
- [15] β arises, in general as a suitable 'large-size' limit of $d \log \sigma_B(E)/dE$. For details on how the passage from microcanonical to canonical, which we have illustrated in detail in the text only for power-law energy-bath densities of states, may be generalized to a wider class of energy-bath models, we refer again to [9] and to [8].
- [16] The fact that both the traditional microcanonical approach, and also the more modern approach based on a total pure state along the lines of Goldstein, Lebowitz et al. and Popescu, only explain how a small part of a totem can be thermal also raises obvious puzzles, not only for black holes, but also for cosmology.
- [17] B.S. Kay, *Entropy defined, entropy increase and decoherence understood, and some black-hole puzzles solved*, arXiv:hep-th/9802172
- [18] B.S. Kay, Class. Quant. Grav. **15**, L89 (1998) [arXiv:hep-th/9810077]
- [19] B.S. Kay and V. Abyaneh, *Expectation values, experimental predictions, events and entropy in quantum gravitationally decohered quantum mechanics*, arXiv:0710.0992
- [20] S.W. Hawking, Phys. Rev. D **14**, 2460 (1976)
- [21] Our proposition in Section ID is not quite a proposition in the sense of pure mathematics and, indeed, it is, in some respects (deliberately) vague. In particular, the phrase "as far as physical quantities of interest are concerned" needs clarifying. What we do expect (and give evidence for in Part 2) is that we may (with high probability and to a very good approximation) at least replace the actual entropy $-\text{tr}(\rho_S^{\text{modern}} \log \rho_S^{\text{modern}})$ by $-\text{tr}(\rho_S^{\text{modapprox}} \log \rho_S^{\text{modapprox}})$ and the moments of energy – i.e. $\bar{\epsilon}^n = \text{tr}(\rho_S^{\text{modern}} H_S^n)$ by $\text{tr}(\rho_S^{\text{modapprox}} H_S^n)$ ($= \int_0^E \epsilon^n P_S(\epsilon) d\epsilon = \text{tr}(\rho_S^{\text{microc}} H_S^n)$ – cf. Equation 33 in Section II). It remains a significant challenge to remove the vagueness and to provide a rigorous proof. However, we expect that such a rigorous proof could be based on the arguments we give in Part 2.
- [22] B.S. Kay, *Modern foundations for thermodynamics and the stringy limit of black hole equilibria* [to appear on arXiv simultaneously with this paper]
- [23] B.S. Kay, *More about the stringy limit of black hole equilibria* [to appear on arXiv simultaneously with this paper]
- [24] Once our 'system', 'S', and 'energy-bath', 'B', are of comparable size, these notions of course become interchangeable and we might alternatively, e.g., call them 'subsystem A' and 'subsystem B'. However we retain the system and energy-bath terminology and notation to maintain contact with the literature in which S is assumed smaller than B. One drawback is that we use the same letter for both system and entropy but we do make them distinguishable by reserving a roman font (S) for system and italic (*S*) for entropy.
- [25] For a discussion of how our proposal differs from the currently standard 'modern' approach (as we are calling it here) see especially Endnote (xii) in [19], where detailed arguments are given for our claim that our proposal offers an understanding of black hole entropy and also of cosmological entropy – while the 'modern' approach by itself, as it would usually be understood, doesn't. In our view, to put thermodynamics on a proper foundation, it is necessary to adopt *both* what we call here the 'modern' approach as well as our matter-gravity entanglement hypothesis. Without the latter, it makes no sense, for example, to ascribe to the whole universe any other entropy value than zero!
- [26] W. Feller, *An Introduction to Probability Theory and its Applications. Volume I* Third Edition (Wiley, New York 1968)
- [27] H. Bauer, *Probability Theory and Elements of Measure Theory* 2nd English edition (Academic, London 1981)
- [28] Our Gaussian approximation (28) to $b(k; n, p)$ (24), is not the same as the more usual approximation [26] which relates the Binomial and Gaussian distributions but it has an amusing relationship to it. The latter, more usual, approximation is, when (for simplicity) we again assume that pn is an integer, the statement

$$b(pn + \kappa; n, p) \simeq \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{\kappa^2}{2np(1-p)}\right). \quad (119)$$

To understand the relationship between (28) and (119) consider, for example the case where $n = 100$ and $p = \frac{1}{2}$. Then (119) with, say, $\kappa = 2$ tells us that if one tosses a fair coin a hundred times, then the probability of getting 52 heads and 48 tails (or vice versa) is approximately $\exp(-2/25)$ times the probability of getting 50 heads and 50 tails. On the other hand, (28) with $x = \pm 1/50$ tells us that the same quantity (i.e. $\exp(-2/25)$ times the probability of getting 50 heads and 50 tails when tossing a fair coin a hundred times) is also an approximation to the probability of getting exactly 50 heads

and 50 tails when one tosses a hundred times a biased coin which has a probability 52/100 of landing heads on each throw (or of landing tails on each throw). The fact that these two distinct probabilities are approximately equal is, arguably, intuitively obvious.

- [29] In fact, allowing different values of c (c_S and c_B) in the obvious places in (34) will, provided c_S and c_B remain of comparable size, only introduce small corrections to the quantities we calculate. In particular, as may be readily checked. (Hint: write $c_S = ce^\mu$ and $c_B = ce^{-\mu}$. Then E_c will be $E/2 - \mu$) the expected energies of S and B will remain exactly unchanged as $E/2$, while there will be only small corrections to the entropies. In particular, S_S^{modern} generalizes from $bE/4 + \log(cE)$ to $bE/4 + \log(c_S c_B E^2)/2 - (\log(c_S/c_B))^2/4E$. Different values of b will, as might be expected, have a more drastic effect. See also [23] where we will consider densities of states of the form $\sigma(\epsilon) = c\epsilon^{-p}e^{b\epsilon}$ – i.e. with inverse-power prefactors such as occur in string theory.
- [30] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series and Products* Corrected and enlarged edition (Academic, London 2007)
- [31] A. Wehrl, Rev. Mod. Phys. **50**, 221 (1978)
- [32] B.S. Kay, Helvetica Physica Acta **58**, 1030 (1985)
- [33] R.T. Powers and E. Størmer, Commun. Math. Phys. **16**, 1 (1970)
- [34] Y. Takahashi and H. Umezawa, Collective Phenomena **2**, 55 (1975)
- [35] A. Mann, M. Revzen and H. Umezawa, Phys. Lett. **139A** 197 (1989)
- [36] B. Yurke and M. Potasek, Phys. Rev. A **36**, 3464 (1987)
- [37] B.S. Kay, Commun. Math. Phys. **100**, 57 (1985)
- [38] L. Susskind, *Some speculations about black hole entropy in string theory*, arXiv:hep-th/9309145
- [39] A. Strominger and C. Vafa, Phys. Lett. **379B**, 99 (1996) [arXiv:hep-th/9601029]
- [40] G.T. Horowitz and J. Polchinski, Phys. Rev. D **55**, 6189 (1997)
- [41] G.T. Horowitz, *Quantum states of black holes*. In R.M. Wald (editor) *Black Holes and Relativistic Stars* (University of Chicago Press, Chicago 1998) [arXiv:gr-qc/9704072]
- [42] In fact we have $M = EK^2\Delta e^{qE^2}(1/qE^2 + O(1/q^2E^4))$ whereupon $P_S(\epsilon) = E^{-1}(qE^2 + O(1))\exp(-2q\epsilon(E - \epsilon))$. This may easily be seen by a direct calculation of M as given by (9) with (62), on using the well-known asymptotic formula $\int_{-a}^a \exp(x^2)dx = \exp(a^2)(1/a + O(1/a^3))$.
- [43] S. W. Hawking, Phys. Rev. D **13**, 191 (1976)
- [44] In the microcanonical ensemble considered in [43], a ‘black hole’ with assumed density of states $\sim \text{const} \exp(4\pi G\epsilon^2)$ is weakly coupled to ‘gravitons’ (but one could add ‘photons’ and include many further matter fields etc.) with a density of states $\sim \text{const} \exp(c\epsilon^{3/4})$ (c a suitable constant). Direct comparison with the microcanonical ensemble we consider here in Section VI (with the replacements, $B \leftrightarrow$ ‘gravity’, $S \leftrightarrow$ ‘matter’) seems not to be straightforward. In particular, it seems that perhaps we should identify Hawking’s ‘black hole’ density of states with the density of states, not for ‘gravity’ alone but rather for what we would regard as being, in his model (extended to many matter fields), ‘gravity’ together with ‘the matter degrees of freedom which are localized near the black hole horizon’.
- [45] Before we can take the logarithm of the long-string density of states, we would need to multiply it by a constant with the dimensions of energy so that the argument of the logarithm is dimensionless. Our criticism is that the Horowitz Polchinski theory does not seem to naturally supply any such constant. It might be countered that our own work does not single out a preferred value of Δ . But this would not be a comparably serious criticism: The role of Δ in our work is only to help us to make an ‘educated guess’ as to the totem state Ψ by choosing a Ψ at random from our energy band $[E, E + \Delta]$. That the choice of Δ is ‘harmless’ is confirmed by the fact (see Section IV) that our formulae for mean energy and entropy are independent of Δ .
- [46] D.N. Page, Phys. Rev. Lett. **71**, 1291 (1993) [arXiv:gr-qc/9305007]
- [47] Page also conjectured the exact general formula $\langle S(\rho_m^{mn}) \rangle = \sum_{k=n+1}^{mn} \frac{1}{k} - \frac{m-1}{2n}$, the correctness of which was, soon afterwards, proven by a number of authors (S.K. Foong and S. Kanno, Phys. Rev. Lett. **72** 1148 (1994); J. Sánchez-Ruiz, Phys. Rev. E **52**, 5653 (1995); S. Sen, Phys. Rev. Lett. **77**, 1 (1996)). What will be of relevance for the present paper however, and particularly for the results we obtain in the remaining sections of Part 2 are mainly just the original results due to Lubkin – in particular, his result on the validity of the approximations (85) and (in Section XIII) what we call there the ‘Lubkin-Page’ approximation (87).
- [48] The equality of $S(\rho_m^{mn})$ and $S(\rho_n^{mn})$ follows by the well-known general result that, for a pure total state on any bipartite system, the von Neumann entropies of the two reduced density operators are necessarily always equal. As is well-known, this is easily proven from the Schmidt decomposition which, by the way, we shall recall, in passing, in another context, in Section XI. In fact, by (107), given an arbitrary state vector, $\Psi \in \mathcal{H}_m \otimes \mathcal{H}_n$ (\mathcal{H}_m and \mathcal{H}_n any m - and n -dimensional Hilbert spaces) we clearly have that the reduced density operator, ρ_m^{mn} , of $|\Psi\rangle\langle\Psi|$ on \mathcal{H}_m is

$$\sum_{i=1}^{\nu} \lambda_i |\tilde{e}_i\rangle\langle\tilde{e}_i|,$$

while the reduced density operator, ρ_n^{mn} , of $|\Psi\rangle\langle\Psi|$ on \mathcal{H}_n is

$$\sum_{i=1}^{\nu} \lambda_i |\tilde{f}_i\rangle\langle\tilde{f}_i|,$$

where $\nu = \min(m, n)$. Clearly, the von Neumann entropy of each of these reduced density operators is $-\sum_{i=1}^{\nu} \lambda_i \log \lambda_i$.

- [49] I. Bengtsson and K. Zyczkowski, *Geometry of Quantum States: An Introduction to Quantum Entanglement* (Cambridge University Press, Cambridge UK 2006)
- [50] G. 't Hooft, Nucl. Phys. **B256**, 727 (1985)
- [51] S. Mukohyama and W. Israel, Phys. Rev. D **58** 104005 (1998)
- [52] B.S. Kay and L. Ortiz, *Brick Walls and AdS/CFT*, arXiv:1111.6429